美国数学会经典影印系列



A Course in Metric Geometry

度量几何学教程

Dmitri Burago
Yuri Burago
Sergei Ivanov

高等教育出版社



A Course in Metric Geometry

度量几何学教程

Dmitri Burago
Yuri Burago
Sergei Ivanov

高等教育出版社·北京

图字：01-2016-2527 号

度量几何学教程
Duliang Jihexue Jiaocheng

图书在版编目(CIP)数据

度量几何学教程 = A Course in Metric Geometry：英文 /（俄罗斯）德米特里・布拉戈 (Dmitri Burago)，（俄罗斯）尤里・布拉戈 (Yuri Burago)，（俄罗斯）谢尔盖・伊万诺夫 (Sergei Ivanov) 著．一影印本．一北京：高等教育出版社，2017.1
ISBN 978-7-04-046908-0
Ⅰ．①度… Ⅱ．①德… ②尤… ③谢… Ⅲ．①几何学一教材一英文 Ⅳ．①O18
中国版本图书馆 CIP 数据核字 (2016) 第 280453 号

策划编辑　赵天夫　　责任编辑　赵天夫
封面设计　张申申　　责任印制　韩　刚

出版发行　高等教育出版社
社址　北京市西城区德外大街 4 号
邮政编码　100120
购书热线　010-58581118
咨询电话　400-810-0598
网址　http://www.hep.edu.cn
http://www.hep.com.cn
网上订购　http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
印刷　涿州市星河印刷有限公司

开本　787mm × 1092mm　1/16
印张　27.25
字数　610 千字
版次　2017 年 1 月第 1 版
印次　2017 年 1 月第 1 次印刷
定价　169.00 元

[物 料 号　46908-00]



美国数学会经典影印系列

出版者的话

近年来，我国的科学技术取得了长足进步，特别是在数学等自然科学基础领域不断涌现出一流的研究成果。与此同时，国内的科研队伍与国外的交流合作也越来越密切，越来越多的科研工作者可以熟练地阅读英文文献，并在国际顶级期刊发表英文学术文章，在国外出版社出版英文学术著作。

然而，在国内阅读海外原版英文图书仍不是非常便捷。一方面，这些原版图书主要集中在科技、教育比较发达的大中城市的大型综合图书馆以及科研院所的资料室中，普通读者借阅不甚容易；另一方面，原版书价格昂贵，动辄上百美元，购买也很不方便。这极大地限制了科技工作者对于国外先进科学技术知识的获取，间接阻碍了我国科技的发展。

高等教育出版社本着植根教育、弘扬学术的宗旨服务我国广大科技和教育工作者，同美国数学会（American Mathematical Society）合作，在征求海内外众多专家学者意见的基础上，精选该学会近年出版的数十种专业著作，组织出版了“美国数学会经典影印系列”丛书。美国数学会创建于 1888 年，是国际上极具影响力的专业学术组织，目前拥有近 30000 会员和 580 余个机构成员，出版图书 3500 多种，冯·诺依曼、莱夫谢茨、陶哲轩等世界级数学大家都是其作者。本影印系列涵盖了代数、几何、分析、方程、拓扑、概率、动力系统等所有主要数学分支以及新近发展的数学主题。

我们希望这套书的出版，能够对国内的科研工作者、教育工作者以及青年学生起到重要的学术引领作用，也希望今后能有更多的海外优秀英文著作被介绍到中国。

高等教育出版社

2016 年 12 月

Preface

This book is not a research monograph or a reference book (although research interests of the authors influenced it a lot)—this is a textbook. Its structure is similar to that of a graduate course. A graduate course usually begins with a course description, and so do we.

Course description. The objective of this book is twofold. First of all, we wanted to give a detailed exposition of basic notions and techniques in the theory of length spaces, a theory which experienced a very fast development in the past few decades and penetrated into many other mathematical disciplines (such as Group Theory, Dynamical Systems, and Partial Differential Equations). However, we have a wider goal of giving an elementary introduction into a broad variety of the most geometrical topics in geometry—the ones related to the notion of distance. This is the reason why we included metric introductions to Riemannian and hyperbolic geometries. This book tends to work with "easy-to-touch" mathematical objects by means of "easy-to-visualize" methods. There is a remarkable book [**Gro3**], which gives a vast panorama of "geometrical mathematics from a metric viewpoint". Unfortunately, Gromov's book seems hardly accessible to graduate students and non-experts in geometry. One of the objectives of this book is to bridge the gap between students and researchers interested in metric geometry, and modern mathematical literature.

Prerequisite. It is minimal. We set a challenging goal of making the core part of the book accessible to first-year graduate students. Our expectations of the reader's background gradually grow as we move further in the book. We tried to introduce and illustrate most of new concepts and methods by using their simplest case and avoiding technicalities that take attention

away from the gist of the matter. For instance, our introduction to Riemannian geometry begins with metrics on planar regions, and we even avoid the notion of a manifold. Of course, manifolds do show up in more advanced sections. Some exercises and remarks assume more mathematical background than the rest of our exposition; they are optional, and a reader unfamiliar with some notions can just ignore them. For instance, solid background in differential geometry of curves and surfaces in $\mathbb{R}^3$ is not a mandatory prerequisite for this book. However, we would hope that the reader possesses some knowledge of differential geometry, and from time to time we draw analogies from or suggest exercises based on it. We also make a special emphasis on motivations and visualizations. A reader not interested in them will be able to skip certain sections. The first chapter is a clinic in metric topology; we recommend that the reader with a reasonable idea of metric spaces just skip it and use it for reference: it may be boring to read it. The last chapters are more advanced and dry than the first four.

Figures. There are several figures in the book, which are added just to make it look nicer. If we included all necessary figures, there would be at least five of them for each page.

• It is a must that the reader systematically studying this book makes a figure for every proposition, theorem, and construction!

Exercises. Exercises form a vital part of our exposition. This does not mean that the reader should solve all the exercises; it is very individual. The difficulty of exercises varies from trivial to rather tricky, and their importance goes all the way up from funny examples to statements that are extensively used later in the book. This is often indicated in the text. It is a very helpful strategy to perceive *every* proposition and theorem as an exercise. You should try to prove each on your own, possibly after having a brief glance at our argument to get a hint. Just reading our proof is the last resort.

Optional material. Our exposition can be conditionally subdivided into two parts: core material and optional sections. Some sections and chapters are preceded by a brief plan, which can be used as a guide through them. It is usually a good idea to begin with a first reading, skipping all optional sections (and even the less important parts of the core ones). Of course, this approach often requires going back and looking for important notions that were accidentally missed. A first reading can give a general picture of the theory, helping to separate its core and give a good idea of its logic. Then the reader goes through the book again, transforming theoretical knowledge into the genuine one by filling it with all the details, digressions, examples and experience that makes knowledge practical.

About metric geometry. Whereas the borderlines between mathematical disciplines are very conditional, geometry historically began from very "down-to-earth" notions (even literally). However, for most of the last century it was a common belief that "geometry of manifolds" basically boiled down to "analysis on manifolds". Geometric methods heavily relied on differential machinery, as it can be guessed even from the name "Differential geometry". It is now understood that a tremendous part of geometry essentially belongs to metric geometry, and the differential apparatus can be used just to define some class of objects and extract the starting data to feed into the synthetic methods. This certainly cannot be applied to all geometric notions. Even the curvature tensor remains an obscure monster, and the geometric meaning of only some of its simplest appearances (such as the sectional curvature) are more or less understood. Many modern results involving more advanced structures still sound quite analytical. On the other hand, expelling analytical machinery from a certain sphere of definitions and arguments brought several major benefits. First of all, it enhanced mathematical understanding of classical objects (such as smooth Riemannian manifolds) both ideologically, and by concrete results. From a methodological viewpoint, it is important to understand what assumptions a particular result relies on; for instance, in this respect it is more satisfying to know that geometrical properties of positively curved manifolds are based on a certain inequality on distances between quadruples of points rather than on some properties of the curvature tensor. This is very similar to two ways of thinking about convex functions. One can say that a function is convex if its second derivative is nonnegative (notice that the definition already assumes that the function is smooth, leaving out such functions as $f(x) = |x|$). An alternative definition says that a function is convex if its epigraph (the set $\{(x, y) : y \geq f(x)\}$) is; the latter definition is equivalent to Jensen's inequality $f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y)$ for all nonnegative α, β with $\alpha + \beta = 1$, and it is robust and does not rely on the notion of a limit. From this viewpoint, the condition $f'' \geq 0$ can be regarded as a convenient criterion for a smooth function to be convex.

As a more specific illustration of an advantage of this way of thinking, imagine that one wants to estimate a certain quantity over all metrics on a sphere. It is so tempting to study a metric for which the quantity attains its maximum, but alas this metric may fail exist within smooth metrics, or even metrics that induce the same topology. It turns out that it still may exist if we widen our search to a class of more general length spaces. Furthermore, mathematical topics whose study used to lie outside the range of noticeable applications of geometrical technique now turned out to be traditional objects of methods originally rooted in differential geometry. Combinatorial group theory can serve as a model example of this

situation. By now the scope of the theory of length spaces has grown quite far from its cradle (which was a theory of convex surfaces), including most of classical Riemannian geometry and many areas beyond it. At the same time, geometry of length spaces perhaps remains one of the most "hands-on" mathematical techniques. This combination of reasons urged us to write this "beginners' course in geometry from a length structure viewpoint".

Acknowledgements. The authors enjoyed hospitality and excellent working conditions during their stays at various institutions, including the University of Strasbourg, ETH Zurich, and Cambridge University. These unforgettable visits were of tremendous help to the progress of this book. The authors' research, which had essential impact on the book, was partially supported by the NSF Foundation, the Sloan Research Fellowship, CRDF, RFBR, and Shapiro Fund at Penn State, whose help we gratefully acknowledge. The authors are grateful to many people for their help and encouragement. We want to especially thank M. Gromov for provoking us to write this book; S. Alexander, R. Bishop, and C. Croke for undertaking immense labor of thoroughly reading the manuscript—their numerous corrections, suggestions, and remarks were of invaluable help; S. Buyalo for many useful comments and suggestions for Chapter 9; K. Shemyak for preparing most of the figures; and finally a group of graduate students at Penn State who took a Math 597c course using our manuscript as the base text and corrected dozens of typos and small errors (though we are confident that twice as many of them are still left for the reader).

Contents

Chapter 1

Metric Spaces

The purpose of the major part of the chapter is to set up notation and to refresh the reader's knowledge of metric spaces and related topics in point-set topology. Section 1.7 contains minimal information about Hausdorff measure and dimension.

It may be a good idea to skip this chapter and use it only for reference, or to look through it briefly to make sure that all examples are clear and exercises are obvious.

1.1. Definitions

Definition 1.1.1. Let X be an arbitrary set. A function $d : X \times X \to \mathbb{R} \cup \{\infty\}$ is a *metric* on X if the following conditions are satisfied for all $x, y, z \in X$.

(1) Positiveness: $d(x, y) > 0$ if $x \neq y$, and $d(x, x) = 0$.

(2) Symmetry: $d(x, y) = d(y, x)$.

(3) Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A *metric space* is a set with a metric on it. In a formal language, a metric space is a pair (X, d) where d is a metric on X. Elements of X are called *points* of the metric space; $d(x, y)$ is referred to as the *distance* between points x and y.

When the metric in question is clear from the context, we also denote the distance between x and y by $|xy|$.

Unless different metrics on the same set X are considered, we will omit an explicit reference to the metric and write "a metric space X" instead of "a metric space (X, d)."

In most textbooks, the notion of a metric space is slightly narrower than our definition: traditionally one consider metrics with finite distance between points. If it is important for a particular consideration that d takes only finite values, this will be specified by saying that d is a *finite metric*. There is a very simple relation between finite and infinite metrics, namely a metric space with possibly infinite distances splits canonically into subspaces that carry finite metrics and are separated from one another by infinite distances:

Exercise 1.1.2. Show that the relation $d(x,y) \neq \infty$ is an equivalence relation. Each of its equivalence classes together with the restriction of d is a metric space with a finite metric.

Definition 1.1.3. Let X and Y be two metric spaces. A map $f : X \to Y$ is called *distance-preserving* if $|f(x)f(y)| = |xy|$ for any two points $x, y \in X$. A bijective distance-preserving map is called an *isometry*. Two spaces are *isometric* if there exists an isometry from one to the other.

It is clear that being isometric is an equivalence relation. Isometric spaces share all properties that can be expressed completely in terms of distances.

Semi-metrics.

Definition 1.1.4. A function $d : X \times X \to \mathbb{R}_+ \cup \{+\infty\}$ is called a *semi-metric* if it satisfies all properties from Definition 1.1.1 of a metric except the requirement that $d(x,y) = 0$ implies $x = y$. This means that we allow zero distance between different points.

There is an obvious relation between semi-metrics and metrics, namely identifying points with zero distance in a semi-metric leads to a usual metric:

Proposition 1.1.5. *Let d be a semi-metric on X. Introduce an equivalence relation R_d on X: set xR_dy iff $d(x,y) = 0$. Since $d(x,y) = d(x_1,y_1)$ whenever xR_dx_1 and yR_dy_1, the projection $\hat{d}$ of d onto the quotient space $\hat{X} = X/R_d$ is well-defined. Then $(\hat{X}, \hat{d})$ is a metric space.*

Proof. Trivial (exercise). □

We will often abuse notation, writing $(X/d, d)$ rather than $(X/R_d, \hat{d})$, with X/d instead of X/R_d and using the same letter d for its projection $\hat{d}$.

Example 1.1.6. Let the distance between two points (x,y), (x',y') in $\mathbb{R}^2$ be defined by $d((x,y),(x',y')) = |(x-x') + (y-y')|$. Check that it is a semi-metric. Prove that the quotient space $(\mathbb{R}^2/d, d)$ is isometric to the real line.

1.2. Examples

Various examples of metric spaces will appear everywhere in the course. In this section we only describe several important ones to begin with. For many of them, verification of the properties from Definition 1.1.1 is trivial and is left for the reader.

Example 1.2.1. One can define a metric on an arbitrary set X by

$$|xy| = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

This example is not particularly interesting but it can serve as the initial point for many constructions.

Example 1.2.2. The real line, $\mathbb{R}$, is canonically equipped with the distance $|xy| = |x - y|$, and thus can be considered as a metric space. There is an immense variety of other metrics on $\mathbb{R}$; for instance, consider $d_{\log}(x, y) = \log(|x - y| + 1)$.

Example 1.2.3. The Euclidean plane, $\mathbb{R}^2$, with its standard distance, is another familiar metric space. The distance can be expressed by the Pythagorean formula,

$$|xy| = |x - y| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

where (x_1, x_2) and (y_1, y_2) are coordinates of points x and y. The triangle inequality for this metric is known from elementary Euclidean geometry. Alternatively, it can be derived from the Cauchy inequality.

Example 1.2.4 (direct products). Let X and Y be two metric spaces. We define a metric on their direct product $X \times Y$ by the formula

$$|(x_1, y_1)(x_2, y_2)| = \sqrt{|x_1x_2|^2 + |y_1y_2|^2}.$$

In particular, $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

Exercise 1.2.5. Derive the triangle inequality for direct products from the triangle inequality on the Euclidean plane.

Example 1.2.6. Recall that the coordinate n-space $\mathbb{R}^n$ is the vector space of all n-tuples $(x_1, \dots, x_n)$ of real numbers, with component-wise addition and multiplication by scalars. It is naturally identified with the multiple direct product $\mathbb{R} \times \cdots \times \mathbb{R}$ (n times). This defines the standard Euclidean distance,

$$|xy| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$.

Example 1.2.7 (dilated spaces). This simple construction is similar to obtaining one set from another by means of a homothety map. Let X be a metric space and $\lambda > 0$. The metric space λX is the same set X equipped with another distance function $d_{\lambda X}$ which is defined by $d_{\lambda X}(x, y) = d_X(x, y)$ for all $x, y \in X$, where d_X is the distance in X. The space λX is referred to as X *dilated* (or *rescaled*) by λ.

Example 1.2.8 (subspaces). If X is a metric space and Y is a subset of X, then a metric on Y can be obtained by simply restricting the metric from X. In other words, the distance between points of Y is equal to the distance between the same points in X.

Restricting the distance is the simplest but not the only way to define a metric on a subset. In many cases it is more natural to consider an *intrinsic metric*, which is generally not equal to the one restricted from the ambient space. The notion of intrinsic metric will be explained further in the course, but its intuitive meaning can be illustrated by the following example of the intrinsic metric on a circle.

Example 1.2.9. The unit circle, S^1, is the set of points in the plane lying at distance 1 from the origin. Being a subset of the plane, the circle carries the restricted Euclidean metric on it. We define an alternative metric by setting the distance between two points as the length of the shorter arc between them. For example, the arc-length distance between two opposite points of the circle is equal to π. The distance between adjacent vertices of a regular n-gon (inscribed into the circle) is equal to $2\pi/n$.

Exercise 1.2.10. (a) Prove that any circle arc of length less or equal to π, equipped with the above metric, is isometric to a straight line segment.

(b) Prove that the entire circle with this metric is not isometric to any subset of the plane (regarded with the restriction of Euclidean distance onto this subset).

1.2.1. Normed vector spaces.

Definition 1.2.11. Let V be a vector space. A function $|\cdot| : V \to \mathbb{R}$ is a *norm* on V if the following conditions are satisfied for all $v, w \in V$ and $k \in \mathbb{R}$.

(1) Positiveness: $|v| > 0$ if $v \neq 0$, and $|0| = 0$.
(2) Positive homogeneity: $|kv| = |k||v|$.
(3) Subadditivity (triangle inequality): $|v + w| \leq |v| + |w|$.

A *normed space* is a vector space with a norm on it. Finite-dimensional normed spaces are also called *Minkowski spaces*. The distance in a normed

space $(V, |\cdot|)$ is defined by the formula

$$d(v, w) = |v - w|.$$

It is easy to see that a normed space with the above distance is a metric space. The norm is recovered from the metric as the distance from the origin.

The Euclidean space $\mathbb{R}^n$ described in Example 1.2.6 is a normed space whose norm is expressed by

$$|(x_1, \dots, x_n)| = \sqrt{x_1^2 + \dots + x_n^2}.$$

There are other natural norms in $\mathbb{R}^n$.

Example 1.2.12. The space $\mathbb{R}_1^n$ is the coordinate space $\mathbb{R}^n$ with a norm $\|\cdot\|_1$ defined by

$$\|(x_1, \dots, x_n)\|_1 = |x_1| + \dots + |x_n|$$

(where $|\cdot|$ is just the absolute value of real numbers).

Example 1.2.13. Similarly, the space $\mathbb{R}_\infty^n$ is $\mathbb{R}^n$ with a norm $\|\cdot\|_\infty$ where

$$\|(x_1, \dots, x_n)\|_\infty = \max\{|x_1|, \dots, |x_n|\}.$$

Exercise 1.2.14. Prove that

(a) $\mathbb{R}_1^2$ and $\mathbb{R}_\infty^2$ are isometric;

(b) $\mathbb{R}_1^n$ and $\mathbb{R}_\infty^n$ are not isometric for any $n > 2$.

Example 1.2.15. Let X be an arbitrary set. The space $\ell_\infty(X)$ is the set of all bounded functions $f : X \to \mathbb{R}$. This is naturally a vector space with respect to pointwise addition and multiplication by scalars. The standard norm $\|\cdot\|_\infty$ on $\ell_\infty(X)$ is defined by

$$\|f\|_\infty = \sup_{x \in X} |f(x)|.$$

Exercise 1.2.16. Show that $\mathbb{R}_\infty^n = \ell_\infty(X)$ for a suitable set X. *Hint*: an n-tuple $(x_1, \dots, x_n)$ is formally a map, isn't it?

1.2.2. Euclidean spaces. Let X be a vector space. Recall that a *bilinear form* on X is a map $F : X \times X \to \mathbb{R}$ which is linear in both arguments. A bilinear form F is *symmetric* if $F(x, y) = F(y, x)$ for all $x, y \in X$. A symmetric bilinear form F can be recovered from its associated *quadratic form* $Q(x) = Q_F(x) = F(x, x)$, e.g., by means of the formula $4F(x, y) = Q(x + y) - Q(x - y)$.

Definition 1.2.17. A *scalar product* is a symmetric bilinear form F whose associated quadratic form is positive definite, i.e., $F(x, x) > 0$ for all $x \neq 0$. A *Euclidean space* is a vector space with a scalar product on it.

We will use notation $\langle \cdot, \cdot \rangle$ for various scalar products.

Definition 1.2.18. A *norm associated with a scalar product* $\langle \cdot, \cdot \rangle$ is defined by the formula $|v| = \sqrt{\langle v, v \rangle}$. A norm is called *Euclidean* if it is associated with some scalar product.

For example, the standard norm in $\mathbb{R}^n$ is associated with the scalar product defined by $\langle x, y \rangle = \sum x_i y_i$ where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$.

Exercise 1.2.19. Prove the triangle inequality for a norm associated with a scalar product.

Hint: First, reduce the triangle inequality to: $\langle v, w \rangle \le |v| \cdot |w|$ for any two vectors v and w. Then expand the relation $\langle v - tw, v - tw \rangle \ge 0$ and substitute $t = \langle v, v \rangle / \langle w, w \rangle$. Another way to prove the triangle inequality is to combine Proposition 1.2.22 and the triangle inequality for $\mathbb{R}^n$.

Since a scalar product is uniquely determined by its associated norm, a Euclidean space could be defined as a normed space whose norm is Euclidean. The following exercise give an explicit characterization of Euclidean spaces among the normed spaces.

Exercise 1.2.20. Prove that a norm $|\cdot|$ on a vector space V is Euclidean if and only if

$$|v + w|^2 + |v - w|^2 = 2(|v|^2 + |w|^2)$$

for all $v, w \in V$.

Exercise 1.2.21. Show that $\mathbb{R}^n_1$ and $\mathbb{R}^n_\infty$ are not Euclidean spaces for $n > 1$.

Two vectors in a Euclidean space are called *orthogonal* if their scalar product is zero. An *orthonormal frame* is a collection of mutually orthogonal unit vectors. Vectors of an orthonormal frame are linearly independent (prove this!). An orthonormal frame can be obtained from any collection of linearly independent vectors by a standard Gram–Schmidt orthogonalization procedure.

In particular, a finite-dimensional Euclidean space V possesses an orthonormal basis. Let $\dim V = n$ and $\{e_1, \dots, e_n\}$ be such a basis. Every vector $x \in V$ can be uniquely represented as a linear combination $\sum x_i e_i$ for some $x_i \in \mathbb{R}$. Since all scalar products of vectors e_i are known, we can find the scalar product of any linear combination, namely

$$\left\langle \sum x_i e_i, \sum y_i e_i \right\rangle = \sum x_i y_i.$$

This implies the following

Proposition 1.2.22. *Every n-dimensional Euclidean space is isomorphic to $\mathbb{R}^n$. This means that there is a linear isomorphism $f : \mathbb{R}^n \to V$ such that $\langle f(x), f(y)\rangle = \langle x, y\rangle$ for all $x, y \in \mathbb{R}^n$. In particular, these spaces are isometric.*

Proof. Define $f((x_1, \dots, x_n)) = \sum x_i e_i$ where $\{e_i\}$ is an orthonormal basis. □

This proposition allows one to apply elementary Euclidean geometry to general Euclidean spaces. For example, since any two-dimensional subspace of a Euclidean space is isomorphic to $\mathbb{R}^2$, any statement involving only two vectors and their linear combinations can be automatically transferred from the standard Euclidean plane to all Euclidean spaces.

Exercise 1.2.23. Prove that any distance-preserving map from one Euclidean space to another is an affine map, that is, a composition of a linear map and a parallel translation. Show by example that this is generally not true for arbitrary normed spaces.

Exercise 1.2.24. Let V be a finite-dimensional normed space. Prove that V is Euclidean if and only if for any two vectors $v, w \in V$ such that $|v| = |w|$ there exists a linear isometry $f : V \to V$ such that $f(v) = w$.

1.2.3. Spheres.

Example 1.2.25. The n-sphere S^n is the set of unit vectors in $\mathbb{R}^{n+1}$, i.e., $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$. The angular metric on S^n is defined by

$$d(x, y) = \arccos \langle x, y\rangle .$$

In other words, the spherical distance is defined as the Euclidean angle between unit vectors. It equals the length of the shorter arc of a great circle connecting x and y in the sphere. Another formula for this metric is

$$d(x, y) = 2 \arcsin \frac{|x - y|}{2}.$$

The metric on the circle described in Example 1.2.9 is a partial case of this example.

1.3. Metrics and Topology

Definition 1.3.1. Let X be a metric space, $x \in X$ and $r > 0$. The set formed by the points at distance less than r from x is called an (open metric) *ball* of radius r centered at x. We denote this ball by $B_r(x)$. Similarly, a *closed ball* $\overline{B}_r(x)$ is the set of points whose distances from x are less than or equal to r.

Exercise 1.3.2. Let x_1 and x_2 be points of some metric space, and let r_1 and r_2 be positive numbers. Show that

(a) if $|x_1x_2| \geq r_1 + r_2$, then the balls $B_{r_1}(x_1)$ and $B_{r_2}(x_2)$ are disjoint;

(b) if $|x_1x_2| \leq r_1 - r_2$, then $B_{r_2}(x_2) \subset B_{r_1}(x_1)$;

(c) the converse statements to (a) and (b) are not always true (give counterexamples).

The topology associated with a metric is defined as follows: a set U in the metric space is open if and only if for every point $x \in U$ there exists an $\varepsilon > 0$ such that $B_\varepsilon(r) \subset U$.

It is easy to see that an open ball is an open set and a closed ball is a closed set (i.e., its complement is open). As a consequence of the former, a set is open if and only if is representable as a union of (possibly infinitely many) open balls.

Exercise 1.3.3. Let X be a metric space and $Y \subset X$. Prove that two topologies on Y coincide: the one associated with the metric restricted on Y, and the subspace topology induced by the one of X (in which a set is open in Y if and only if it is representable as an intersection of Y and an open set in X).

Exercise 1.3.4. Prove that a metric product carries the standard product topology.

Definition 1.3.5. A sequence $\{x_n\}_{n=1}^\infty$ of points of a topological space X is said to *converge* to a point $x \in X$ if for any neighborhood U of x there is a number n_0 such that $x_n \in U$ for all $n \geq n_0$. Notation: $x_n \to x$ (as $n \to \infty$). The point x is called a *limit* of the sequence.

In a metric space, $x_n \to x$ if and only if $|x_nx| \to 0$. The following properties are also specific for metric spaces.

Proposition 1.3.6. *Let X and Y be metric spaces. Then*

(1) *A sequence in X cannot have more than one limit.*

(2) *A point $x \in X$ is an accumulation point of a set $S \subset X$ (i.e., belongs to the closure of S) if and only if there exists a sequence $\{x_n\}_{n=1}^\infty$ such that $x_n \in S$ for all n and $x_n \to x$. In particular, S is closed if and only if it contains all limits of sequences contained within S.*

(3) *A map $f : X \to Y$ is continuous at a point $x \in X$ if and only if $f(x_n) \to f(x)$ for any sequence $\{x_n\}$ converging to x.*

1.4. Lipschitz Maps

Definition 1.4.1. Let X and Y be metric spaces. A map $f : X \to Y$ is called *Lipschitz* if there exists a $C \geq 0$ such that $|f(x_1)f(x_2)| \leq C|x_1x_2|$ for all $x_1, x_2 \in X$. Any suitable value of C is referred to as a *Lipschitz constant* of f. The minimal Lipschitz constant is called the *dilatation* of f and denoted by $\operatorname{dil} f$. The dilatation of a non-Lipschitz function is infinity.

A map with Lipschitz constant 1 is called *nonexpanding.*

Exercise 1.4.2. The distance from a point x to a set S in a metric space is defined by $\operatorname{dist}(x, S) = \inf_{y \in S} |xy|$. Prove that $\operatorname{dist}(\cdot, S)$ is a nonexpanding function.

Proposition 1.4.3. (1) *All Lipschitz maps are continuous.*

(2) *If $f : X \to Y$ and $g : Y \to Z$ are Lipschitz maps, then $g \circ f$ is Lipschitz and $\operatorname{dil}(g \circ f) \leq \operatorname{dil} f \cdot \operatorname{dil} g$.*

(3) *The set of real-valued Lipschitz functions on a metric space (and, more generally, the set of Lipschitz functions from a metric space to a normed space) is a vector space. One has $\operatorname{dil}(f+g) \leq \operatorname{dil} f + \operatorname{dil} g$, $\operatorname{dil}(\lambda f) = |\lambda| \operatorname{dil} f$ for any Lipschitz functions f and g and $\lambda \in \mathbb{R}$.*

Definition 1.4.4. Let X and Y be metric spaces. A map $f : X \to Y$ is called *locally Lipschitz* if every point $x \in X$ has a neighborhood U such that $f|_U$ is Lipschitz. The *dilatation of f at x* is defined by

$$\operatorname{dil}_x f = \inf\{\operatorname{dil} f|_U : U \text{ is a neighborhood of } x\}.$$

Exercise 1.4.5. Let X be a metric space. Prove that $\operatorname{dil} f = \sup_{x \in \mathbb{R}} \operatorname{dil}_x f$ for any map $f : \mathbb{R} \to X$. Prove the same statement with $\mathbb{R}$ replaced by S^1 with the metric described in Exercise 1.2.9. Show that it is not true for S^1 with the metric restricted from $\mathbb{R}^2$.

Definition 1.4.6. Let X and Y be metric space. A map $f : X \to Y$ is called *bi-Lipschitz* if there are positive constants c and C such that

$$c|x_1x_2| \leq |f(x_1)f(x_2)| \leq C|x_1x_2|$$

for all $x_1, x_2 \in X$.

Clearly every bi-Lipschitz map is a homeomorphism onto its image.

Definition 1.4.7. Two metrics d_1 and d_2 on the same set X are called *bi-Lipschitz* if there are positive constants c and C such that

$$c \cdot d_1(x, y) \leq d_2(x, y) \leq C \cdot d_1(x, y)$$

for all $x, y \in X$.

In other words, d_1 and d_2 are Lipschitz equivalent if the identity is a bi-Lipschitz map from (X, d_1) to (X, d_2). Clearly this is an equivalence relation on the set of metrics in X. Lipschitz equivalent metrics determine the same topology.

Exercise 1.4.8. Give an example of two metrics on the same set that determine the same topology but are not Lipschitz equivalent.

Exercise 1.4.9. Let X and Y be metric spaces. Prove that the following three metrics on $X \times Y$ are Lipschitz equivalent:

1. The metric defined in Example 1.2.4.
2. $d_1((x_1, y_1), (x_2, y_2)) = |x_1x_2| + |y_1y_2|$.
3. $d_\infty((x_1, y_1), (x_2, y_2)) = \max\{|x_1x_2|, |y_1y_2|\}$.

Exercise 1.4.10. Let X be a metric space. Prove that its metric is a Lipschitz function on $X \times X$ where $X \times X$ is regarded to have any of the metrics from the previous exercise.

We conclude this section with the following important theorem about normed spaces.

Theorem 1.4.11. *1. Two norms on a vector space determine the same topology if and only if they are Lipschitz equivalent;*

2. All norms on a finite-dimensional vector space are Lipschitz equivalent.

1.5. Complete Spaces

Definition 1.5.1. A sequence $\{x_n\}$ in a metric space is called a *Cauchy sequence* if $|x_nx_m| \to 0$ as $n, m \to \infty$. The precise meaning of this is the following: for any $\varepsilon > 0$ there exists an n_0 such that $|x_nx_m| < \varepsilon$ whenever $n \geq n_0$ and $m \geq n_0$.

A metric space is called *complete* if every Cauchy sequence in it has a limit.

It is known from analysis (see e.g. [**Mun**]) that $\mathbb{R}$ is a complete space. It easily follows that $\mathbb{R}^n$ is complete for all n. $\mathbb{R} \setminus \{0\}$ is an example of a noncomplete space; a sequence that would converge to zero in $\mathbb{R}$ is a Cauchy sequence that has no limit in this space. (Note that a converging sequence is always a Cauchy one.)

Exercise 1.5.2. Prove that completeness is preserved by a bi-Lipschitz homeomorphism. In particular, Lipschitz equivalent metrics share completeness or noncompleteness.

Exercise 1.5.3. Show that completeness is *not* a topological property; i.e., there exist homeomorphic metric spaces X and Y such that X is complete but Y is not.

Exercise 1.5.4. The *diameter* of a set S in a metric space is defined by $\operatorname{diam}(S) = \sup_{x,y \in S} |xy|$. Prove that a metric space X is complete if and only if it possesses the following property. If $\{X_n\}$ is a sequence of closed subsets of X such that $X_{n+1} \subset X_n$ for all n, and $\operatorname{diam}(X_n) \to 0$ as $n \to 0$, then the sets X_n have a common point.

Show that the assumption $\operatorname{diam}(X_n) \to 0$ is essential.

Proposition 1.5.5. *Let X be a metric space and $Y \subset X$. Then*

(1) *If Y is complete, then Y is closed in X.*

(2) *If X is complete and Y is closed in X, then Y is complete.*

The following two exercises provide useful tools for proving completeness of some spaces.

Exercise 1.5.6. Let $\{x_n\}$ be a Cauchy sequence in a metric space. Prove that

(a) If $\{x_n\}$ has a converging subsequence, then it converges itself.

(b) For any sequence $\{\varepsilon_n\}$ of positive numbers there exists a subsequence $\{y_n\}$ of $\{x_n\}$ such that $|y_n y_{n+1}| < \varepsilon_n$ for all n.

Exercise 1.5.7. Let $\{x_n\}_{n=1}^{\infty}$ be a sequence in a metric space such that the series $\sum_{n=1}^{\infty} |x_n x_{n+1}|$ has a finite sum. Prove that $\{x_n\}$ is a Cauchy sequence.

Exercise 1.5.8 (fixed-point theorem)**.** Let X be a complete space, $0 < \lambda < 1$, and let $f : X \to X$ be a map such that $|f(x)f(y)| \le \lambda |xy|$ for all $x, y \in X$. Prove that there exists a unique point $x_0 \in X$ such that $f(x_0) = x_0$.

Hint: Obtain x_0 as the limit of a sequence $\{x_n\}$ where x_1 is an arbitrary point and $x_{n+1} = f(x_n)$ for all $n \ge 1$.

The following simple proposition is used many times in this book.

Proposition 1.5.9. *Let X be a metric space and X' a dense subset of X. Let Y be a complete space and $f : X' \to Y$ a Lipschitz map. Then there exists a unique continuous map $\tilde{f} : X \to Y$ such that $\tilde{f}|_{X'} = f$. Moreover $\tilde{f}$ is Lipschitz and* $\operatorname{dil} \tilde{f} = \operatorname{dil} f$.

Proof. Let C be a Lipschitz constant for f. For every $x \in X$ define $\tilde{f}(x) \in Y$ as follows. Choose a sequence $\{x_n\}_{n=1}^{\infty}$ such that $x_n \in X'$ for all n, and $x_n \to x$ as $n \to \infty$. Observe that $\{f(x_n)\}$ is a Cauchy sequence in Y. Indeed, we have $|f(x_i)f(x_j)| \le C|x_i x_j|$ for all i, j, and $|x_i x_j| \to 0$

as $i, j \to \infty$ because the sequence $\{x_n\}$ converges. Therefore the sequence $\{f(x_n)\}$ converges; then define $\tilde{f}(x) = \lim_{n\to\infty} f(x_n)$.

Thus we have defined a map $\tilde{f} : X \to Y$. Then the inequality $|\tilde{f}(x)\tilde{f}(x')| \le C|xx'|$ for $x, x' \in X$ follows as a limit of similar inequalities for f. Indeed, if $x = \lim x_n$, $x' = \lim x'_n$, $\tilde{f}(x) = \lim f(x_n)$, $\tilde{f}(x') = \lim f(x'_n)$, then

$$|\tilde{f}(x)\tilde{f}(x')| = \lim_{n\to\infty} |f(x_n)f(x'_n)| \le C \lim_{n\to\infty} |x_n x'_n| = C|xy|.$$

Therefore f is Lipschitz (and hence continuous) and $\operatorname{dil} f \le C$.

The uniqueness of $\tilde{f}$ is trivial: if two continuous maps coincide on a dense set, then they coincide everywhere. □

Completion. Inside a metric space there is an operation of taking closure that makes a closed subset out of an arbitrary subset. The following theorem defines a similar operation that makes a complete metric space out of a noncomplete one.

Theorem 1.5.10. *Let X be a metric space. Then there exists a complete metric space $\tilde{X}$ such that X is a dense subspace of $\tilde{X}$. It is essentially unique in the following sense: if $\tilde{X}'$ is another space with these properties, then there exists a unique isometry $f : \tilde{X} \to \tilde{X}'$ such that $f|_X = \mathrm{id}$.*

Definition 1.5.11. The space $\tilde{X}$ from the above theorem is called the *completion* of X.

Proof of Theorem 1.5.10. Let $\mathfrak{X}$ denote the set of all Cauchy sequences in X. Introduce the distance in $\mathfrak{X}$ by the formula

$$d(\{x_n\}, \{y_n\}) = \lim_{n\to\infty} |x_n y_n|.$$

It is easy to check that, if $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences, then $\{|x_n y_n|\}$ is either a Cauchy sequence of real numbers or $|x_n y_n| = \infty$ for all large enough n. Therefore the above limit always exists. Clearly d is a semi-metric on $\mathfrak{X}$. Define $\tilde{X} = X/d$ (see Proposition 1.1.5 and a remark after it).

There is a natural map from X to $\tilde{X}$, namely let a point $x \in X$ be mapped to a point of $\tilde{X}$ represented by the constant sequence $\{x\}_{n=1}^{\infty}$. Since this map is distance-preserving, we can identify X with its image in $\tilde{X}$ (formally, change the definition of $\tilde{X}$ so that points of X replace their images). This way X becomes a subset of $\tilde{X}$. It is dense because a point of $\tilde{X}$ represented by a sequence $\{x_n\}$ is the limit of this sequence (thought of as the sequence in $X \subset \tilde{X}$).

The uniqueness part of the theorem follows from Proposition 1.5.9 applied to the inclusion maps from X to $\tilde{X}$ and $\tilde{X}'$. □

Baire's theorem.

Definition 1.5.12. A set Y in a topological space X is *nowhere dense* if the closure of Y has empty interior.

Equivalently, Y is nowhere dense in X if the interior of $X \setminus Y$ is dense. By plugging in the definitions of closure and interior, one obtains the following description: Y is nowhere dense if and only if any open set U contains a ball which does not intersect Y.

Theorem 1.5.13 (Baire's theorem). *A complete metric space cannot be covered by countably many nowhere dense subsets. Moreover, a union of countably many nowhere dense subsets has a dense complement.*

Remark 1.5.14. An equivalent formulation is: in a complete space, an intersection of countably many sets whose interiors are dense (in particular, an intersection of countably many open dense sets) is dense.

Remark 1.5.15. A union of countably many nowhere dense sets may not be nowhere dense. For example, consider $\mathbb{Q} \subset \mathbb{R}$ as a union of single points.

Proof of the theorem. Let X be a complete metric space and $\{Y_i\}_{i=1}^{\infty}$ be a countable family of nowhere dense sets. We have to show that any open set $U \subset X$ contains a point which does not belong to $\bigcup_{i=1}^{\infty} Y_i$. Since Y_1 is nowhere dense, there is a (closed) ball $B_1 \subset U$ which does not intersect Y_1. Since Y_2 is nowhere dense, there is a closed ball $B_2 \subset B_1$ which does not intersect Y_2. And so on. This way we obtain a sequence $B_1 \supset B_2 \supset \dots$ of closed balls where each ball B_i has no common points with the respective set Y_i. We may choose the radii of the balls B_i so that they converge to zero. Then the centers of the balls form a Cauchy sequence. The limit of this sequence belongs to all balls and therefore does not belong to any of the sets Y_i. □

1.6. Compact Spaces

Recall that a topological space X is called *compact* if any open covering of X (that is, a collection of open sets that cover X) has a finite sub-collection that still covers X. The term "compact set" refers to a subset of a topological space that is compact with respect to its induced topology.

Definition 1.6.1. Let X be a metric space and $\varepsilon > 0$. A set $S \subset X$ is called an *ε-net* if $\operatorname{dist}(x, S) \le \varepsilon$ for every $x \in X$.

X is called *totally bounded* if for any ε there is a finite ε-net in X.

Exercise 1.6.2. Let X be a metric space, $Y \subset X$ and $\varepsilon > 0$. A set $S \in X$ is called an *ε-net for* Y if $\operatorname{dist}(y, S) \le \varepsilon$ for all $y \in Y$. Prove that, if there is an ε-net for Y, then there exists a (2ε)-net for Y contained in Y.

Exercise 1.6.3. Prove that

(a) Any subset of a totally bounded set is totally bounded.

(b) In $\mathbb{R}^n$, any bounded set (that is, a set whose diameter is finite) is totally bounded.

Exercise 1.6.4. A set S in a metric space is called *ε-separated,* for an $\varepsilon > 0$, if $|xy| \geq \varepsilon$ for any two different points $x, y \in S$. Prove that

1 If there exists an $(\varepsilon/3)$-net of cardinality n, then an ε-separated set cannot contain more than n points.

2. A maximal ε-separated set is an ε-net.

The following theorem gives a list of equivalent definitions of compactness for metric spaces. The last one is the most important for us.

Theorem 1.6.5. *Let X be a metric space. Then the following statements are equivalent:*

(1) *X is compact.*
(2) *Any sequence in X has a converging subsequence.*
(3) *Any infinite subset of X has an accumulation point.*
(4) *X is complete and totally bounded.*

The following properties are known from general topology.

Proposition 1.6.6. *Let X and Y be Hausdorff topological spaces. Then*

(1) *If $S \subset X$ is a compact set, then S is closed in X.*
(2) *If X is compact and $S \subset X$ is closed in X, then S is compact.*
(3) *If $\{X_n\}_{n=1}^{\infty}$ is a sequence of compact sets such that $X_{n+1} \subset X_n$ for all n, then the $\bigcap_{n=1}^{\infty} X_n \neq \emptyset$.*
(4) *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*
(5) *If X is compact and $f : X \to Y$ is a continuous map, then $f(X)$ is a compact set.*
(6) *If X is compact and $f : X \to Y$ is bijective continuous map, then f is a homeomorphism.*
(7) *If X is compact and $f : X \to \mathbb{R}$ is a continuous function, then f attains its maximum and minimum.*

The property of $\mathbb{R}^n$ expressed in the fourth statement is known as *boundedly compactness.*

Definition 1.6.7. A metric space is said to be *boundedly compact* if all closed bounded sets in it are compact.

Exercise 1.6.8. Prove that a metric space (with possibly infinite distances) is compact if and only if it is a union of a finite number of compact subsets each of which carries a finite metric.

Exercise 1.6.9. Let X be a compact metric space. Prove that

1. If the metric of X is finite, then $\operatorname{diam} X < \infty$.
2. There exist two points $x, y \in X$ such that $|xy| = \operatorname{diam} X$.

Exercise 1.6.10. Define the distance between two subsets A and B of a metric space X by $\operatorname{dist}(A, B) = \inf\{|xy| : x \in A, y \in B\}$. (Warning: this kind of distance does not satisfy triangle inequality!) Prove that

1. If A and B are compact, then there exist $x \in A$ and $y \in B$ such that $|xy| = \operatorname{dist}(A, B)$.

2. If $X = \mathbb{R}^n$, the same is true under the weaker assumptions that A is compact and B is closed.

Theorem 1.6.11 (Lebesgue's Lemma). *Let X be a compact metric space, and let $\{U_\alpha\}_{\alpha \in A}$ be an open covering of X. Then there exists a $\rho > 0$ such that any ball of radius ρ in X is contained in one of the sets U_α.*

Proof. We may assume that the metric of X is finite and none of the sets U_α covers the whole space. Then one can define a function $f : X \to \mathbb{R}$ by

$$f(x) = \sup\{r \in \mathbb{R} : B_r(x) \text{ is contained in one of the } U_\alpha\}.$$

Since $\{U_\alpha\}$ is an open covering, $f(x)$ is well-defined and positive for all $x \in X$. Clearly f is a nonexpanding function and hence continuous. Therefore it attains a (positive) minimum r_0. Define $\rho = r_0/2$. □

The number ρ from the theorem is referred to as a *Lebesgue number* of the covering.

Theorem 1.6.12. *Let X and Y be metric spaces and let X be compact. Then every continuous map $f : X \to Y$ is uniformly continuous, i.e., for every $\varepsilon > 0$ there is a $\delta > 0$ such that for all $x_1, x_2 \in X$ such that $|x_1x_2| < \delta$ one has $|f(x_1)f(x_2)| < \varepsilon$.*

Proof. Every $x \in X$ has a neighborhood U such that $f(U) \subset B_{\varepsilon/2}(f(x))$, in particular, $\operatorname{diam}(f(U)) < \varepsilon$. Hence open sets U such that $\operatorname{diam}(f(U)) < \varepsilon$ cover X. Let δ be a Lebesgue number of this covering. □

Exercise 1.6.13. Prove that a locally Lipschitz map from a compact space is Lipschitz.

Isometries of compact spaces. Unlike most of this chapter (which is rather analytical), the following two theorems have purely geometric contents. The first of them is a simple special case of the second one.

Theorem 1.6.14. *A compact metric space cannot be isometric to a proper subset of itself. In other words, if X is a compact space and $f : X \to X$ is a distance-preserving map, then $f(X) = X$.*

Proof. Suppose the contrary, i.e., let $p \in X \setminus f(X)$. Since $f(X)$ is compact and hence closed, there exists an $\varepsilon > 0$ such that $B_\varepsilon(p) \cap f(X) = \emptyset$. Let n be the maximal possible cardinality of an ε-separated set in X (see Exercise 1.6.4) and let $S \subset X$ be an ε-separated set of cardinality n. Since f is distance-preserving, the set $f(S)$ is also ε-separated. On the other hand, $\operatorname{dist}(p, f(S)) \geq \operatorname{dist}(p, f(X)) \geq \varepsilon$ and therefore $f(S) \cup \{p\}$ is an ε-separated set of cardinality $n + 1$. Contradiction. □

Theorem 1.6.15. *Let X be a compact metric space. Then*

(1) *Any nonexpanding surjective map $f : X \to X$ is an isometry.*
(2) *If a map $f : X \to X$ is such that $|f(x)f(y)| \geq |xy|$ for all $x, y \in X$, then f is an isometry.*

Proof. 1. Suppose the contrary, i.e., that $|f(p)f(q)| < |pq|$ for some points $p, q \in X$. Fix p and q and pick an $\varepsilon > 0$ such that $|f(p)f(q)| < |pq| - 5\varepsilon$.

Let n be a natural number such that there exists at least one ε-net in X of cardinality n. Consider the set $\mathfrak{N} \subset X^n$ of all n-tuples of points of X that form ε-nets in X. This set is closed in X^n and therefore it is compact. Define a function $D : X^n \to \mathbb{R}$ by

$$D(x_1, \dots, x_n) = \sum_{i,j=1}^{n} |x_i x_j|.$$

This function is continuous and therefore it attains a minimum on $\mathfrak{N}$. Let $S = (x_1, \dots, x_n)$ be an element of $\mathfrak{N}$ at which the minimum is attained. Since f is nonexpanding and surjective, the collection $f(S) := (f(x_1), \dots, f(x_n))$ is also an element of $\mathfrak{N}$. Moreover $D(f(S)) \leq D(S)$ because $|f(x_i)f(x_j)| \leq |x_i x_j|$ for all i, j. But $D(S)$ is the minimum of D on $\mathfrak{N}$; therefore $D(f(S)) = D(S)$ and $|f(x_i)f(x_j)| = |x_i x_j|$ for all i, j.

On the other hand, there exist indices i and j such that $|px_i| \leq \varepsilon$ and $|qx_j| \leq \varepsilon$. For these i and j we have

$$|x_i x_j| \geq |pq| - |px_i| - |qx_j| \geq |pq| - 2\varepsilon$$

and

$$\begin{aligned} |f(x_i)f(x_j)| &\leq |f(p)f(q)| + |f(p)f(x_i)| + |f(q)f(x_j)| \\ &\leq |f(p)f(q)| + 2\varepsilon \leq |pq| - 3\varepsilon. \end{aligned}$$

Hence $|f(x_i)f(x_j)| < |x_ix_j|$. Contradiction.

2. Define $Y = f(X)$. The argument from the proof of the previous theorem shows that Y is dense in X. Consider the map $g = f^{-1} : Y \to X$. Since g is nonexpanding and Y is dense in X, g can be extended to a nonexpanding map $\tilde{g} : X \to X$. By the first part of the theorem, $\tilde{g}$ is an isometry. Hence $Y = X$ and f is an isometry. □

1.7. Hausdorff Measure and Dimension

1.7.1. Measures in general. The notion of measure generalizes length, area and volume. Roughly speaking, a measure on a space X is a nonnegative function defined on a set of subsets of X and possessing the additivity property of the area; namely, the measure of a union of disjoint sets equals the sum of measures of these sets. In fact, a stronger requirement of countable additivity (or σ-additivity, see definitions below) is imposed to make measures really useful.

Although it is commonly accepted that one can speak of area for figures on the plane and volume of three-dimensional bodies, it is not so easy to give correct definitions of these notions. In fact, the statement "every set has a volume" is wrong, as the following fact shows (see **[HM]** for details). Let B denote a unit ball in $\mathbb{R}^3$ with its center removed. Then B can be split into four disjoint subsets, which can be rearranged (by means of rotations) so as to form two copies of B. If volume was defined for sets such as these four pieces (which are in fact extremely wild sets), then B and the union of its disjoint copies would have equal volumes, implying that the volume of all sets is zero. Therefore one has to restrict the class of sets for which the volume (or other measure) is defined. A class of sets on which a measure is defined (these sets are called *measurable*, w.r.t. the measure) must be a σ-algebra, which is defined as follows:

Definition 1.7.1. Let X be an arbitrary set. A set $\mathfrak{A}$ of subsets of X is called a *σ-algebra* if it satisfies the following conditions:

(1) $\emptyset$ and X are elements of $\mathfrak{A}$;

(2) If $A, B \in \mathfrak{A}$, then $A \setminus B \in \mathfrak{A}$;

(3) If $\{A_i\}_{i\in I}$ is a finite or countable collection of elements of $\mathfrak{A}$, then their union $\bigcup_{i\in I} A_i$ is also an element of $\mathfrak{A}$.

Remark 1.7.2. Due to the formula $\bigcup(X \setminus A_i) = X \setminus \bigcap A_i$, a σ-algebra contains any intersection of a countable collection of its elements.

Definition 1.7.3. A *measure* on a σ-algebra $\mathfrak{A}$ is a function $\mu : \mathfrak{A} \to \mathbb{R}_+ \cup \{+\infty\}$ such that

(1) $\mu(\emptyset) = 0$; and

(2) if $\{A_i\}$ is a finite or countable collection of elements of $\mathfrak{A}$ and the sets A_i are disjoint, then $\mu(\bigcup A_i) = \sum \mu(A_i)$.

The second condition is referred to as *σ-additivity.* Note that the expression $\sum \mu(A_i)$ is either a finite sum or a series; its value is well defined and independent of the order of terms since the terms are nonnegative.

Exercise 1.7.4. Let μ be a measure. Prove the following statements:

(a) Let $\{A_i\}_{i=1}^{\infty}$ be a sequence of measurable sets such that $A_i \subset A_{i+1}$ for all i. Then the sequence $\{\mu(A_i)\}$ is nondecreasing and $\lim \mu(A_i) = \mu(\bigcup A_i)$.

(b) Let $\{A_i\}_{i=1}^{\infty}$ be a sequence of measurable sets such that $A_i \supset A_{i+1}$ for all i, and assume that $\mu(A_1) < \infty$. Then the sequence $\{\mu(A_i)\}$ is nonincreasing and $\lim \mu(A_i) = \mu(\bigcap A_i)$.

(c) The assumption $\mu(A_1) < \infty$ in (b) is essential.

If $\mathfrak{S}$ is an arbitrary collection of subsets of a set X, there obviously exists a unique minimal σ-algebra containing $\mathfrak{S}$ (prove this!); it is called the σ-algebra *generated by* $\mathfrak{S}$. If X is a topological space, then the σ-algebra generated by its topology (i.e., by the set of all open sets) is called the *Borel σ-algebra* of X. Elements of the Borel σ-algebra are called *Borel sets.* A measure defined on the Borel σ-algebra is called a *Borel measure* over X.

The following theorem provides a basis for measure theory in Euclidean spaces.

Theorem 1.7.5 (Lebesgue). *There exists a unique Borel measure m_n over $\mathbb{R}^n$ which is invariant under parallel translations and such that $m_n([0,1]^n) = 1$.*

The measure m_n is called *Lebesgue measure.* The uniqueness part of the theorem implies that every translation-invariant Borel measure in $\mathbb{R}^n$ with a finite value for a cube is a constant multiple of m_n.

Exercise 1.7.6. Prove that

1. Lebesgue measure is invariant under isometries of $\mathbb{R}^n$.

2. If $L : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map, then $m_n(L(A)) = |\det L| \cdot m_n(A)$ for any measurable set $A \subset \mathbb{R}^n$. In particular, a homothety with coefficient C multiplies the Lebesgue measure by C^n.

Hint: Use the uniqueness part of Theorem 1.7.5.

1.7.2. Hausdorff measure. To motivate the definition of Hausdorff measures, let us recall the main idea of the construction behind the proof of Theorem 1.7.5. It begins with choosing a class of simple sets (such as balls or cubes). Then a set (from an appropriate σ-algebra) is covered by simple sets; then the Lebesgue measure of the set is defined as the infimum of total

measures of such covers. Speaking about "the total measure of a cover" one means here that certain measure is already assigned to simple sets.

To define Hausdorff n-dimensional measure on a metric space, one could proceed along the same lines: cover a set by metric balls such that all their radii are less than ε. For each ball, consider a Euclidean ball of the same radius and add their volumes for all balls from the cover: this will be the total measure of the cover. Taking its infimum over all covers and passing to the limit as ε approaches zero, one gets a version of Hausdorff measure of the set. Instead of adding volumes of Euclidean balls of the same radii, one could simply add the radii of balls from the cover raised to the power n: the result is the same up to a constant multiplier. It turns out that arbitrary sets and diameters are technically more convenient to use than metric balls and radii.

Now we pass to formal definitions.

Definition 1.7.7. Let X be a metric space and d be a nonnegative real number.

For a finite or countable covering $\{S_i\}_{i\in I}$ of X (that is, a collection of sets such that $X \subset \bigcup S_i$), define its *d-weight* $w_d(\{S_i\})$ by the formula

$$w_d(\{S_i\}) = \sum_i (\operatorname{diam} S_i)^d.$$

If $d = 0$, substitute each (if any) 0^0 term in the formula by 1.

For an $\varepsilon > 0$ define $\mu_{d,\varepsilon}(X)$ by

$$\mu_{d,\varepsilon}(X) = \inf\big\{w_d(\{S_i\}) : \operatorname{diam}(S_i) < \varepsilon \text{ for all } i\big\}.$$

The infimum is taken over all finite or countable coverings of X by sets of diameter $< \varepsilon$; if no such covering exists, then the infimum is $+\infty$.

The *d-dimensional Hausdorff measure of* X is defined by the formula

$$\mu_d(X) = C(d) \cdot \lim_{\varepsilon\to 0} \mu_{d,\varepsilon}(X)$$

where $C(d)$ is a positive normalization constant. This constant is introduced for only one reason: for integer d it is convenient to choose $C(d)$ so that the d-dimensional Hausdorff measure of a unit cube in $\mathbb{R}^n$ equals 1. In fact, almost nothing depends on the actual value of $C(d)$.

It may be unclear from the definition what the value of $\mu_d(\emptyset)$ is. We explicitly define $\mu_d(\emptyset) = 0$ for all $d \geq 0$.

Clearly $\mu_{d,\varepsilon}(X)$ is a nonincreasing function of ε. Since such a function has a (possibly infinite) limit as $\varepsilon \to 0$, $\mu_d(X)$ is well defined for any metric space X. It may be either a nonnegative real number or $+\infty$.

Though we have defined Hausdorff measure for a metric space, this notion will often be applied to subsets of metric spaces. In such cases, a subset should be considered as a metric space with the restricted metric. (Note that the definition can be read verbatim if X is a subset of a larger metric space; it does not matter whether covering sets S_i are actually contained in X.)

The following proposition summarizes the properties of Hausdorff measure that follow immediately from the definition.

Proposition 1.7.8. *Let X and Y be metric spaces, and let A and B be subsets of X. Then*

(1) *If $A \subset B$, then $\mu_d(A) \leq \mu_d(B)$.*
(2) *$\mu_d(\bigcup A_i) \leq \sum \mu_d(A_i)$ for any finite or countable collection of sets $A_i \subset X$.*
(3) *If $\operatorname{dist}(A, B) > 0$, then $\mu_d(A \cup B) = \mu_d(A) + \mu_d(B)$.*
(4) *If $f : X \to Y$ is a Lipschitz map with a Lipschitz constant C, then $\mu_d(f(X)) \leq C^d \cdot \mu_d(X)$.*
(5) *If $f : X \to Y$ is a C-homothety, i.e., $|f(x_1)f(x_2)| = C|c_1x_2|$ for all $x_1, x_2 \in X$, then $\mu_d(f(X)) = C^d \cdot \mu_d(X)$.*

According to Carathéodory's criterion, ([**Fe**], 2.3.1(9)), any nonnegative function on the Borel σ-algebra of X possessing the properties 1–3 from Proposition 1.7.8 is actually a measure. Thus we obtain

Theorem 1.7.9. *For any metric space X and any $d \geq 0$, μ_d is a measure on the Borel σ-algebra of X.*

Exercise 1.7.10. Prove that 0-dimensional Hausdorff measure of a set is its cardinality. In other words, $\mu_0(X)$ is a number of points in X if X is a finite set, and $\mu_0(X) = \infty$ if X is an infinite set.

Exercise 1.7.11. Let X and Y be metric spaces and $f : X \to Y$ a locally Lipschitz map with dilatation $\leq C$. Prove that $\mu_d(f(X)) \leq C^d \cdot \mu_d(X)$ assuming that (a) X is compact; (b) X has a countable topological base.

1.7.3. Hausdorff measure in $\mathbb{R}^n$. Let I denote the interval $[0, 1]$ of $\mathbb{R}$. Then $I^n = [0, 1]^n$ is the unit cube in $\mathbb{R}^n$.

Theorem 1.7.12. $0 < \mu_n(I^n) < \infty$.

Exercise 1.7.13. Prove the theorem.

Now we can define the normalization constant $C(n)$ from the definition of Hausdorff measure. Namely, choose $C(n)$ so that $\mu_n(I^n) = 1$. The existence of such a constant follows from Theorem 1.7.12. Theorem 1.7.5 then implies

that Hausdorff measure μ_n on $\mathbb{R}^n$ coincides with the standard n-dimensional volume m_n.

In most cases, the actual value of $C(n)$ is not important. However it is an interesting fact that $C(n)$ equals the volume of the Euclidean n-ball of diameter 1. The proof is based on the following theorem.

Theorem 1.7.14 (Vitali's Covering Theorem). *Let X be a bounded set in $\mathbb{R}^n$ and let $\mathfrak{B}$ be a collection of closed balls in $\mathbb{R}^n$ such that for every $x \in X$ and $\varepsilon > 0$ there is a ball $B \in \mathfrak{B}$ such that $x \in B$ and $\operatorname{diam}(B) < \varepsilon$. Then $\mathfrak{B}$ contains a finite or countable subcollection $\{B_i\}$ of disjoint balls which covers X up to a set of zero measure, i.e., such that $B_i \cap B_j = \emptyset$ if $i \neq j$ and $\mu_n(X \setminus \bigcup_i B_i) = 0$.*

Proof. We may assume that every ball $B \in \mathfrak{B}$ contains at least one point of X and exclude the balls with radius greater than 1. Then all these balls are contained in the 2-neighborhood of X which is bounded and hence has finite volume. We construct a sequence $\{B_i\}_{i=1}^{\infty}$ of balls by induction. If $B_1, \dots, B_m$ are already constructed, we choose the next ball B_{m+1} as follows. Let $\mathfrak{B}_m$ denote the set of balls from the collection that do not intersect any of $B_1, \dots, B_m$. If $\mathfrak{B}_m$ is empty, then $B_1 \cup \cdots \cup B_m$ covers the entire set X and the proof is finished (this follows from the condition that every point is covered by balls of arbitrarily small radii). If $\mathfrak{B}_m$ is not empty, choose B_{m+1} to be any element of $\mathfrak{B}_m$ with

$$\operatorname{diam}(B_{m+1}) > \frac{1}{2} \sup\{\operatorname{diam}(B) : B \in \mathfrak{B}_m\}. \tag{1.1}$$

The balls B_i are disjoint by the construction. We will now show that they cover X up to a set of zero measure. Fix an $\varepsilon > 0$. Since the balls are disjoint and are contained in a set of finite volume, we have $\sum_{i=0}^{\infty} \mu_n(B_i) < \infty$. Hence there is an index m such that $\sum_{i=m+1}^{\infty} \mu_n(B_i) < \varepsilon$. Let $x \in X \setminus \bigcup_i B_i$ and let B be any ball from the collection that contains x and does not intersect the balls $B_1, \dots, B_m$. Note that B must intersect $\bigcup_i B_i$ because otherwise $B \in \mathfrak{B}_m$ for all m which contradicts that $\mu_n(B_i) \to 0$. Let k be the minimal index such that $B \cap B_k \neq \emptyset$. Then $B \in \mathfrak{B}_{m-1}$ and hence $\operatorname{diam}(B_k) > \frac{1}{2} \operatorname{diam}(B)$ by (1.1). It follows that the distance from x to the center of B_k is not greater than 5 times the radius of B_k. Hence x belongs to the ball with the same center as B_k and radius 5 times larger. We denote this ball by $5B_k$.

We have just proved that every $x \in X \setminus \bigcup_i B_i$ belongs to a ball $5B_k$ for some $k > m$. Thus $X \setminus \bigcup_i B_i \sup \bigcup_{i=m+1}^{\infty} (5B_i)$; hence

$$\mu_n(X \setminus \bigcup_i B_i) \leq \sum_{i=m+1}^{\infty} \mu_n(5B_i) = 5^n \sum_{i=m+1}^{\infty} \mu_n(B_i) < 5^n \varepsilon.$$

Since ε is arbitrary, it follows that $\mu_n(X \setminus \bigcup_i B_i) = 0$. □

Corollary 1.7.15. *The normalization constant for the n-dimensional Hausdorff measure equals the volume of the Euclidean n-ball of diameter 1.*

Proof. Let C_n denote the constant from the formulation. Then the volume of a Euclidean n-ball equals $C_n d^n$ where d is its diameter. Let μ'_n be the n-dimensional Hausdorff measure with normalization constant C_n. We have to prove that $\mu'_n = \mu_n$, i.e., that $\mu'_n(I^n) = 1$.

1. $\mu'_n(I^n) \le 1$. To prove this, apply Theorem 1.7.14 to the set $\mathfrak{B}$ of all closed Euclidean balls contained in I^n. This yields a countable collection $\{B_i\}$ of such balls such that the set $Y = I^n \setminus \bigcup B_i$ has zero measure. Hence $\mu'_n(I^n) \le \mu'_n(Y) + \sum C_n \operatorname{diam}(B_i)^n = 0 + \sum m_n(B_i) \le m_n(I^n) = 1$.

2. $\mu'_n(I^n) \ge 1$. By a well-known Bieberbach inequality (cf. e.g. [**BZ**], Theorem 11.2.1), a Euclidean ball has the maximal volume among the sets with the same diameter. Hence $m_n(S) \le C_n \operatorname{diam}(S)^n$ for any bounded set $S \subset \mathbb{R}^n$. Now if $\{S_i\}_{i=1}^{\infty}$ is a covering of I^n, then $1 = m_n(I^n) \le \sum m_n(S_i) \le \sum C_n \operatorname{diam}(S)^n$. The statement follows. □

1.7.4. Hausdorff dimension. The next theorem tells us how the Hausdorff measure of a fixed set depends on dimension. Briefly, the measure is zero or infinite for all dimensions except at most one. More precisely, there is a "critical dimension" below which the measure is infinity and above which the measure is zero. This dimension is an important characteristic of a metric space, called the Hausdorff dimension. Warning: at the critical dimension, all three possibilities (the measure is zero, positive number or $+\infty$) may take place.

Theorem 1.7.16. *For a metric space X there exists a $d_0 \in [0, +\infty]$ such that $\mu_d(X) = 0$ for all $d > d_0$ and $\mu_d(X) = \infty$ for all $d < d_0$.*

Proof. Define $d_0 = \inf\{d \ge 0 : \mu_d(X) \ne \infty\}$. Trivially $\mu_d(X) = \infty$ for all $d < d_0$. If $d > d_0$, there is a $d' < d$ such that $\mu_{d'}(X) = M < \infty$. Therefore for any $\varepsilon > 0$ there exists a covering $\{S_i\}$ of X such that $\operatorname{diam} S_i < \varepsilon$ for all i and $\sum (\operatorname{diam} S_i)^{d'} < 2M$. Then

$$\sum (\operatorname{diam} S_i)^d \le \varepsilon^{d-d'} \cdot \sum (\operatorname{diam} S_i)^{d'} \le 2\varepsilon^{d-d'} M.$$

Hence $\mu_{d,\varepsilon}(X) \le 2\varepsilon^{d'-d} M$. Since $\varepsilon^{d'-d} \to 0$ as $\varepsilon \to 0$, we have $\mu_d(X) = 0$. □

Definition 1.7.17. The value d_0 from Theorem 1.7.16 is called the *Hausdorff dimension* of X and denoted by $\dim_H(X)$.

Remark 1.7.18. Hausdorff dimension is not necessarily integer.

Here are some immediate properties of Hausdorff dimension.

Proposition 1.7.19. *Let X be a metric space. Then*

(1) *If $Y \subset X$, then $\dim_H(Y) \le \dim_H(X)$.*

(2) *If X is covered by a finite or countable collection $\{X_i\}$ of its subsets, then $\dim_H(X) = \sup_i \dim_H(X_i)$.*

(3) *If $f : X \to Y$ is a Lipschitz map, then $\dim_H(f(X)) \le \dim_H(X)$. In particular, bi-Lipschitz equivalent metric spaces have equal Hausdorff dimensions.*

(4) $\dim_H(\mathbb{R}^n) = \dim_H(I^n) = n$.

Exercise 1.7.20. Let X and Y be metric spaces and $f : X \to Y$ a map such that $|f(x_1)f(x_2)| \le C \cdot |x_1x_2|^\alpha$ for all $x_1, x_2 \in X$, where C and α are some positive constants. Prove that $\dim_H(f(X)) \le \dim_H(X)/\alpha$.

Exercise 1.7.21. Prove that the Hausdorff dimension of the standard Cantor set is $\log_3 2$. More generally, let X be a compact space that can be split into n subsets $X_1, \dots, X_n$ that can be obtained from X by dilations with coefficients $c_1, \dots, c_n$ respectively. Prove that $d = \dim_H(X)$ satisfies the equation $\sum c_i^d = 1$. (Warning: compactness is essential!)

Exercise 1.7.22. Give examples of

(a) an uncountable metric space whose Hausdorff dimension is zero;

(b) a metric space X with $\dim_H(X) = 1$ and $\mu_1(X) = 0$;

(c) a metric space X with $\dim_H(X) = 1$ and $\mu_1(X) = +\infty$.

Chapter 2

Length Spaces

2.1. Length Structures

First we want to informally illustrate our main concept. Imagine that you ask a mathematician: "What is the distance between New York and Sydney?". Perhaps, you get the answer "about 8 thousand miles". It is formally correct and still absolutely useless: this is the length of a straight tunnel through the Earth. Analogously, every mountaineer knows that distance in mountains is a tricky thing: if you measure it by an optical device, you get the distance "as a crow flies". It may be relevant for a crow, while wingless creatures confined to the surface of the Earth (like us) have to take long detours with lots of ups and downs; see Figure 2.1.

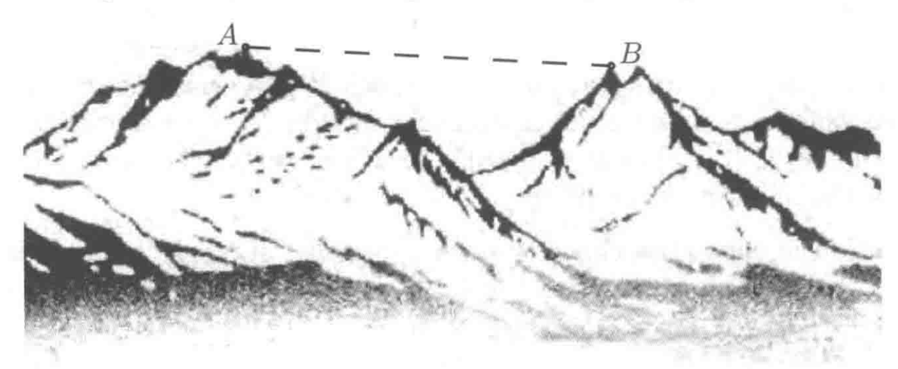


Figure 2.1: "A crow flies" along the segment AB; for a pedestrian it probably takes longer.

This little philosophical digression contains a very clear mathematical moral: in many cases, we have to begin with length of paths as the primary notion and only after that can we derive a distance function. Let us make this observation slightly more precise. For every two points on a surface in

Euclidean space (you may keep thinking of the surface of the Earth) we can measure Euclidean distance between the two points. What we do instead is we introduce a new distance which is measured along the shortest path between the two points. Generalizing this idea, one says that a distance function on a metric space is an intrinsic metric if the distance between two points can be realized by paths connecting the points (mathematically, it must be equal to the infimum of lengths of paths between the points—a shortest path may not exist).

If length of paths is our primary notion, one readily asks for its rigorous definition, where it may arise from and what are the properties of such structures. We will be occupied with these questions throughout this book.

2.1.1. Definition of length structures. Loosely speaking, a length structure consists of a *class of admissible paths* for which we can measure their length, and the *length* itself, which is a correspondence assigning a nonnegative number to every path from the class. Both the class and the correspondence have to possess several natural properties; in all reasonable examples (and in particular in all examples in this book) these requirements are automatically satisfied.

From now on we reserve the word *path* for maps of intervals: a path γ in a (topological) space X is a (continuous) map $\gamma : I \to X$ defined on an interval $I \subset \mathbb{R}$. By an *interval* we mean any connected subset of the real line; it may be open or closed, finite or infinite, and a single point is counted as an interval. Since a path is a map one can speak about its image, restrictions, etc.

A length structure on a topological space X is a class A of admissible paths, which is a subset of all continuous paths in X, together with a map $L : A \to \mathbb{R}_+ \cup \{\infty\}$; the map is called length of path. The class A has to satisfy the following assumptions:

(1) The class A is closed under restrictions: if $\gamma : [a, b] \to X$ is an admissible path and $a \le c \le d \le b$, then the restriction $\gamma|_{[c,d]}$ of γ to $[c, d]$ is also admissible.

(2) A is closed under concatenations (products) of paths. Namely, if a path $\gamma : [a, b] \to X$ is such that its restrictions γ_1,γ_2 to $[a, c]$ and $[c, b]$ are both admissible paths, then so is γ. (Recall that γ is called the *product* or *concatenation* of γ_1 and γ_2, $\gamma = \gamma_1 \cdot \gamma_2$).

(3) A is closed under (at least) linear reparameterizations: for an admissible path $\gamma : [a, b] \to X$ and a homeomorphism $\varphi : [c, d] \to [a, b]$ of the form $\varphi(t) = \alpha t + \beta$, the composition $\gamma \circ \varphi(t) = \gamma(\varphi(t))$ is also an admissible path.

Remark 2.1.1. Every natural class of paths comes with its own class of reparameterizations. For example, consider the class of all continuous paths and the class of homeomorphisms, the class of piecewise smooth paths and the class of diffeomorphisms. We only require that this class of reparameterizations includes all linear maps.

Examples of such classes include: all continuous paths; piecewise smooth paths (on a smooth manifold); broken lines in $\mathbb{R}^n$; see other examples below.

We require that L possesses the following properties:

(1) Length of paths is additive: $L(\gamma_{|_{[a,b]}}) = L(\gamma_{|_{[a,c]}}) + L(\gamma_{|_{[c,b]}})$ for any $c \in [a,b]$.

(2) The length of a piece of a path continuously depends on the piece. More formally, for a path $\gamma : [a,b] \to X$ of finite length, denote by $L(\gamma, a, t)$ the length of the restriction of $\gamma : [a,b] \to X$ to the segment $[a,t]$. We require that $L(\gamma, a, \cdot)$ be a continuous function. (Observe that the previous property implies that $L(\gamma, a, a) = 0$.)

(3) The length is invariant under reparameterizations: $L(\gamma \circ \varphi) = L(\gamma)$ for a linear homeomorphism φ.

(In fact, all reasonable length structures are invariant under arbitrary reparameterizations: $L(\gamma \circ \varphi) = L(\gamma)$ for any homeomorphism φ such that both γ and $\gamma \circ \varphi$ are admissible. However, it is not necessary to verify this in the beginning.)

(4) We require length structures to agree with the topology of X in the following sense: for a neighborhood U_x of a point x, the length of paths connecting x with points of the complement of U_x is separated from zero:

$$\inf\{L(\gamma) : \gamma(a) = x, \gamma(b) \in X \setminus U_x\} > 0.$$

There are several important types of length structures that will appear in this course. When the reader meets with these structures, it is advisable to come back to this definition and make sure that all of them belong to the same general scheme.

Notation. We will often use the notation $L(\gamma, a, b)$ introduced above. Namely, if $\gamma : I \to X$ is an (admissible) path and $[a,b] \subset I$, where $a \le b$, we will denote by $L(\gamma, a, b)$ the length of the restriction of γ to $[a,b]$, i.e., $L(\gamma, a, b) = L(\gamma|_{[a,b]})$. In addition, we define $L(\gamma, b, a) = -L(\gamma, a, b)$. This convention implies that $L(\gamma, a, b) = L(\gamma, a, c) + L(\gamma, c, b)$ for all $a, b, c \in I$ (verify this).

2.1.2. Length spaces. Once we have a length structure, we are ready to define a metric (a distance function) associated with the structure. We will

always assume that the topological space X carrying the length structure is a Hausdorff space. For two points $x, y \in X$ we set the associated distance $d(x,y)$ between them to be the infimum of lengths of admissible paths connecting these points:

$$d_L(x,y) = \inf\{L(\gamma); \gamma : [a,b] \to X, \gamma \in A, \gamma(a) = x, \gamma(b) = y\}.$$

If it is clear from the context which length structure L gives rise to d_L, we usually drop L in the notation d_L.

Exercise 2.1.2. Verify that (X, d_L) is a metric space.

Note that d_L is not necessarily a finite metric. For instance, if X is a disconnected union of two components, no continuous path can go from one component to the other and therefore the distance between points of different components is infinite. On the other hand, there may be points such that continuous paths connecting them exist but all have infinite length. One says that two points $x, y \in X$ belong to the same accessibility component if they can be connected by a path of finite length.

Exercise 2.1.3. 1. Check that accessibility by paths of finite length is indeed an equivalence relation. Your argument should use additivity of length and the assumption that the concatenation of admissible paths is an admissible path.

2. Verify that accessibility components coincide with components of finiteness for d_L.

3. Verify that accessibility components coincide with both connectivity and path connectivity components of (X, d_L).

Did you notice that you have used the following fact?

Exercise 2.1.4. Prove that admissible paths of finite length are continuous with respect to (X, d_L).

This exercise deals with the topology determined by the metric d_L rather than the initial topology of the space X. There really are examples where these two topologies differ; such examples will appear later in this book.

Exercise 2.1.5. Prove that the topology determined by d_L can be only finer than that of X: any open set in X is open in (X, d_L) as well.

Definition 2.1.6. A metric that can be obtained as the distance function associated to a length structure is called an *intrinsic*, or *length, metric*. A metric space whose metric is intrinsic is called a *length space*.

Not every metric can arise as a length metric. Even if (X, d) is a length space and $A \subset X$, the restriction of d to A is not necessarily intrinsic. For example, consider a circle in the plane.

Moreover, not every metrizable topology can be induced by intrinsic metrics:

Exercise 2.1.7. 1. Prove that the set of rational numbers is not homeomorphic to a length space.

2. Prove that the union of the graph $\{(x,y) : y = \sin(1/x),\ x > 0\}$ and the y-axis (with its topology inherited from $\mathbb{R}^2$) is not homeomorphic to a length space.

There can be more delicate reasons why a topological space may be not homeomorphic to a length space:

Exercise 2.1.8. Consider the union of segments

$$\bigcup_{i=1}^{\infty}[(0,0),(\cos 1/i, \sin 1/i)] \cup [(0,0),(1,0)]$$

in the Euclidean plane, depicted in Figure 2.2.

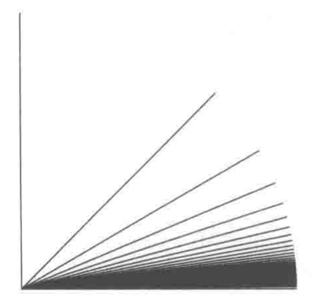


Figure 2.2: The space (with the topology inherited from $\mathbb{R}^2$) is not homeomorphic to a length space.

This set (resembling a fan made of segments) is a topological space with its topology inherited from the Euclidean plane (this is the topology of Euclidean distance restricted to the set). Prove that this topological space is not homeomorphic to a length space.

As a hint to the above exercises, consider the following more general one.

Exercise 2.1.9. Prove that a length space is locally path connected: every neighborhood of any point contains a smaller neighborhood which is path-connected.

One uses infimum instead of simple minimum when defining d_L since there may be no shortest path between two points. For instance, consider

the Euclidean plane with an open segment removed; then for the endpoints of the segment, the shortest path does not exist: it is just removed. Still, its length can be approximated with a given precision by other paths connecting the points. Such situations rarely arise in "real-life" examples; in most cases they will be also prohibited by imposing completeness-compactness type assumptions. On the other hand, existence of shortest paths helps to avoid tedious and nonessential complications. For the simplicity of our exposition we will often restrict ourselves to complete length structures, which are defined as follows:

Definition 2.1.10. A length structure is said to be *complete* if for every two points x, y there exists an admissible path joining them whose length is equal to $d_L(x, y)$; in other words, a length structure is complete if there exists a shortest path between every two points.

Intrinsic metrics associated with complete length structures are said to be *strictly intrinsic*.

2.2. First Examples of Length Structures

To get better motivated, let us briefly meet with a few examples from the zoo of length structures and intrinsic metrics; we will not analyze them in this section, but we suggest that you keep them in mind and use them for testing further definitions and concepts. Notice that in many examples the space itself is a part of a Euclidean space, and there are two different ways of changing the usual Euclidean length structure: we change the class of admissible paths or change the notion of length of paths (or both).

Example 2.2.1 ("Driving in Manhattan"). The space here is the Euclidean plane, and length of paths is the same as usual. The only difference is that we restrict the class of admissible paths to broken lines with edges parallel to one of the coordinate axes. (A critically thinking reader should yell that paths are maps while broken lines are sets! This is absolutely true, and formally we mean the paths whose images are broken lines.) Can you draw a ball in the corresponding intrinsic metric? (It will not look round: you should get a diamond.)

Example 2.2.2 ("Metric on an island"). The space is a connected region in the Euclidean plane, and again length of paths is the same as usual. Admissible paths are all (piecewise smooth) paths contained in the region. If the region is convex, this length structure induces usual Euclidean distance. One may think of this region as an island, and the distance is measured by a creature who cannot swim. Drawing balls in intrinsic metrics arising this way may be quite fun; see Figure 2.3. Is this metric strictly intrinsic? What if we consider the closure of the region?

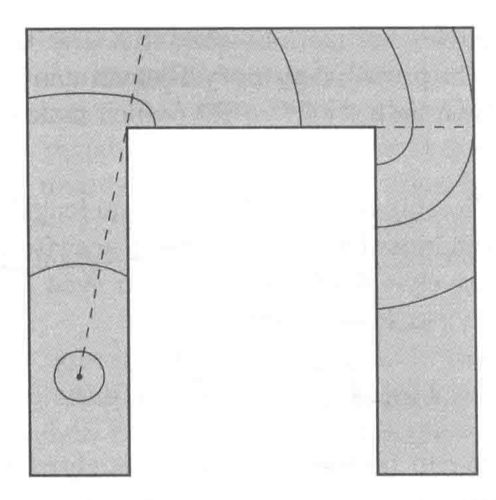


Figure 2.3: Metric balls in a nonconvex island.

The reader can generalize this example for a subspace of a space with length structure. Certainly, only sensible choices for a subspace lead to reasonable examples. For instance, restricting the Euclidean length structure in $\mathbb{R}^2$ to a circle leads to the angular metric. More generally, one obtains spherical geometry by restricting the usual Euclidean length structure to a round sphere. On the other hand, restricting the standard length structure of $\mathbb{R}$ to the set of rational points we obtain a space with each accessibility component consisting of just one point.

Example 2.2.3 (induced length structure). The formal contents of this example is comprised in the following definition. Let $f : X \to Y$ be a continuous map from a topological space X to a space Y endowed with a length structure. One defines the *induced length structure* in X as follows. A path in X is admissible if its composition with f is admissible in Y. The length of an admissible path in X is set to the length of its composition with f with respect to the length structure in Y.

(In fact, this construction may not define a length structure in X because the new length function may fail to satisfy the fourth condition from section 2.1.1. We use the term "induced length structure" only if this is indeed a length structure.)

At first glance, the above definition may sound like a tautology. However, the properties of an induced metric may drastically differ from the properties of a metric we began with. For instance, the leading example of an induced metric when f is a *surface* (that is an immersion $f : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$ of a two-dimensional region into $\mathbb{R}^3$) has served as the main motivating example in metric geometry for over a century. For a reader who is already familiar

with Riemannian metrics, we mention that it is also true (though hard to believe and not easy to prove) that every Riemannian length structure on $\mathbb{R}^n$ can be induced by a map $f : \mathbb{R}^n \to \mathbb{R}^n$ (which makes lots of folds and is rarely smooth).

Example 2.2.4 ("Crossing a swamp": conformal length). The space is the Euclidean plane, and admissible paths are all (piecewise smooth) paths. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a positively-valued continuous (or even L_∞) function. Define the length of a path $\gamma : [a, b] \to \mathbb{R}^2$ by

$$L(\gamma) = \int_a^b f(\gamma(t)) \cdot |\gamma'(t)| dt.$$

This length structure can be thought of as a weighted Euclidean distance. For instance, a traveler who measures the length (=time needed to cover) of a certain route would apparently assign big values to f in a territory that is difficult to traverse (for instance, a swamp or a mountain trail). From the mathematical viewpoint, this is the first example of a Riemannian length structure, which will be discussed further in Chapter 5; the word "conformal" in the title of this subsection reflects the fact that such types of Riemannian structures are called conformally flat.

Example 2.2.5 (Finslerian length). Thinking of the previous example as a length structure for a traveler who assigns weights to different parts of his/her path, one notices that an important feature of real travel is not reflected here. Namely, the difficulty of traversing a region depends not only on the region itself but also on the direction of the route; for instance, choosing a direction in which most ravines are oriented might essentially simplify the trip. To incorporate this additional information, one introduces a function f in two variables and applies it to both γ and its velocity γ'. The expression for the length reads:

$$L(\gamma) = \int_a^b f(\gamma(t), \gamma'(t))\, dt.$$

(A physics-oriented reader recognizes that this structure can be interpreted as action.) In order for this expression to be invariant under bijective reparameterizations of paths, one has to require that f satisfies $f(x, kv) = |k| f(x, v)$ for all scalars k, points x and vectors v (check this as an exercise for change of variable in a definite integral). Usually a stronger requirement is imposed on f, namely for every point x the function $f(x, \cdot)$ must be a norm. A motivation for this will be explained in section 2.4.2. Length structures obtained from this type of constructions are called *Finslerian*, or *Finsler*.

Remark 2.2.6. A reader who seriously tests our definitions against needs of travelers and mountaineers will notice that some features are still missing.

Namely the fact that walking downhill may be easier than climbing uphill cannot be reflected in a length structure. Since the distance in a metric space must be symmetric, we had to require that the length is invariant under all changes of variable including the orientation-reversing ones like $t \mapsto -t$. One could modify the definitions to allow nonsymmetric length structures and metrics. Everything in this chapter can be adapted to such generalized settings; however, this would not make sense in the rest of the book.

Example 2.2.7 (A "cobweb" and a "notebook"). Begin with several disjoint segments and glue some of their endpoints together. Such a space may resemble a cobweb in Euclidean space. This space has a natural length structure. All continuous paths are admissible. The space is built out of segments, and we know how to measure the length of a path while it travels within one segment. Thus to find the length of a path we restrict it to (countably many) intervals such that the image of each interval is contained in one segment, and add the lengths of the restrictions. This is a first example of metric graphs, and the construction of its length structure is a particular case of gluing, discussed in detail in section 3.1.

Another example of the same type can be made out of several copies of a closed half-plane by attaching them together along their boundary lines. This is an example of a polyhedral length space. It can be visualized (and realized) in Euclidean spaces: it looks like an open book. Can you modify the definition of the length structure on a cobweb for this case? Caution: while we could disregard the part of a path spent in endpoints of segments (nodes of the cobweb) since they have zero length, this is not the case for the common edge of the half-planes.

2.3. Length Structures Induced by Metrics

2.3.1. Length of curves in metric spaces. Let us recall our motivating example from the very beginning of this chapter. We began with some distance function (Euclidean distance "as a crow flies") that was not satisfactory since there might be no paths realizing this distance. By saying this we already mean that we know how to measure length of paths, and this length is somehow derived from the Euclidean distance!

Indeed, some of the main examples of length structures are those induced by metric structures. For admissible paths one may use just all continuous paths; for some of them the length may be infinite. In some cases a better choice is the class of Lipschitz paths, that is, the class of maps $\gamma : [a, b] \to X$ such that $d_X(\gamma(t), \gamma(t')) \leq C|t - t'|$, for all $t, t' \in [a, b]$; C is a positive constant.

How do we define the length of a path in Euclidean space? We approximate the path by broken lines and define the length as the limit of their lengths. For each of these broken lines, its vertices belong to (the image of) the path and they are well ordered with respect to the parameter of the path. Since all that we actually use are distances between neighboring vertices in this list, we can mimic this definition in a general metric space in the most straightforward way (compare also with 2.4.12).

Definition 2.3.1. Let (X, d) be a metric space and γ be a path in X, i.e., a continuous map $\gamma : [a, b] \to X$. Consider a *partition* Y of $[a, b]$), that is, a finite collection of points $Y = \{y_0, ..., y_N\}$ such that $a = y_0 \le y_1 \le y_2 \le \cdots \le y_N = b$. The supremum of the sums

$$\Sigma(Y) = \sum_{i=1}^{N} d(\gamma(y_{i-1}), \gamma(y_i)).$$

over all the partitions Y is called the *length* of γ (with respect to the metric d) and denoted $L_d(\gamma)$. A curve is said to be *rectifiable* if its length is finite.

The length structure induced by the metric d is defined as follows: all continuous paths (parameterized by closed intervals) are admissible, and the length is given by the function L_d.

This definitions can be formally applied to any metric space, but one gets sensible examples only by a wise choice of a metric space to begin with: for instance, if we start with a discrete space, there are no nonconstant continuous paths at all. If it is clear from the context which metric d induces the length L, we usually drop d in the notation L_d.

The usual "Euclidean" definition uses passing to a limit as the edges of broken lines approach zero. The following exercise shows that there is no difference here:

Exercise 2.3.2. Prove that $\Sigma(Y) \to L(\gamma)$ as $\max_i\{|y_i - y_{i+1}|\} \to 0$.

Exercise 2.3.3. Prove that the definition of length is compatible with the one used in differential geometry. Namely if $(V, |\cdot|)$ is a finite-dimensional normed vector space and $\gamma : [a, b] \to V$ is a differentiable map, then $L(\gamma) = \int_a^b |\gamma'(t)|\, dt$.

2.3.2. Properties of the induced length. All properties of length structures hold for this structure; this length is also *semi-continuous*. Let us verify (some of) them:

Proposition 2.3.4. *The length structure $L = L_d$ induced by a metric d possesses the following properties:*

(i) *Generalized triangle inequality: $L(\gamma) \ge d(\gamma(a), \gamma(b))$.*

(ii) Additivity: if $a < c < b$, then $L(\gamma, a, c) + L(\gamma, c, b) = L(\gamma)$. In particularly, $L(\gamma, a, c)$ is a nondecreasing function of c.

(iii) If γ is rectifiable, the function $L(\gamma|_{[c,d]}) = L(\gamma, c, d)$ is continuous in c and d.

(iv) L is a lower semi-continuous functional on the space of continuous maps of $[a, b]$ in X with respect to point-wise convergence, and hence with respect to the uniform (i.e., C^0-) topology. This means that if a sequence of rectifiable paths γ_i (with the same domain) is such that $\gamma_i(t)$ converges to $\gamma(t)$ (as $i \to \infty$, for every t in the domain), then $\liminf L(\gamma_i) \geq L(\gamma)$.

Proof. (i) Indeed, the triangle inequality implies that $\Sigma(Y) \geq d(\gamma(a), \gamma(b))$ for all Y's, and thus the inequality persists under passing to the limit.

(ii) First notice that if Y' is obtained from Y by adding one point, then $\Sigma(Y) \geq \Sigma(Y)$ (by the same triangle inequality). Adding c to a partition Y of $[a, b]$ and then splitting it into two partitions of $[a, c]$ and $[c, b]$ completes the argument.

We repeat again that the readers should try to consider such lemmas as exercises and try to prove them on their own. If reading the proof was needed, then drawing a figure with all notations is a must!

(iii) We prove the continuity of L in d, $a < d \leq b$, from the left (the other cases are analogous and are left to the reader). Take $\varepsilon > 0$ and consider a partition Y such that $L(\gamma) - \Sigma(Y) < \varepsilon$. One may suppose that $y_{j-1} < d = y_j$. Then

$$L(\gamma, y_{j-1}, d) - d(\gamma(y_{j-1}), \gamma(d)) < \varepsilon,$$

and the same inequality takes place for each c such that $y_{j-1} \leq c \leq d$.

(iv) Let paths γ_j converge pointwise to γ. Take $\varepsilon > 0$ and fix a partition Y for γ such that $L(\gamma) - \Sigma(Y) < \varepsilon$. Now consider the sums $\Sigma_j(Y)$ for paths γ_j corresponding to the same partition Y. Choose j to be so large that the inequality $d(\gamma_j(y_i), \gamma(y_i)) < \varepsilon$ holds for all $y_i \in Y$. Then

$$L(\gamma) \leq \Sigma(Y) + \varepsilon \leq \Sigma_j(Y) + \varepsilon + (N+1)\varepsilon \leq L(\gamma_j) + (N+2)\varepsilon.$$

Since ε is arbitrary, this implies (iv). □

Remark 2.3.5. In general, functional L is not continuous. A stairs-like example is shown in Figure 2.4.

2.3.3. Induced intrinsic metric. A metric d induces a length structure. The latter, in its turn, gives rise to an intrinsic metric (on each component of accessibility by rectifiable paths). Thus we obtain a canonical construction of induced intrinsic metrics

$$(X, d) \to (X, \widehat{d})$$

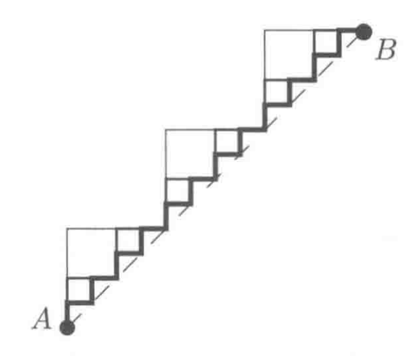


Figure 2.4: L is not continuous.

where $\widehat{d} = d_{L_d}$.

Exercise 2.3.6. Prove that the intrinsic metric induced by the restriction of Euclidean distance to the circle $x^2 + y^2 = 1$ is the angular metric.

Note that the topology of the induced intrinsic metric may be very poorly connected with the original topology of the space, as can be seen from the following examples-exercises:

Exercise 2.3.7. Find the induced intrinsic metric for the metric

$$d((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + \sqrt{|y_1 - y_2|}$$

on $\mathbb{R}^2$. What is the topology of the resulting length space?

Answer: A continuum of disjoint real lines, each with its standard metric.

Exercise 2.3.8. Consider the union of segments

$$U = \bigcup_{n=1}^{\infty} [(0,1), (1/n, 0)] \cup [(0,1), (0,0)] \subset \mathbb{R}^2.$$

The sequence of points $\{(1/n, 0)\}$ converges to $(0,0)$ in the topology inherited by U from $\mathbb{R}^2$. Since all pairwise distances between these points in the induced intrinsic metric are at least 2, this sequence diverges with respect to the intrinsic metric. Prove these statements.

Exercise 2.3.9. Connecting the point $(0,1) \in \mathbb{R}^2$ with all points of the standard Cantor set in the segment $[0,1] = [0,1] \times \{0\} \subset \mathbb{R}^2$, one obtains a compact connected set. Show that in the induced intrinsic metric this set is noncompact although it is still connected.

Exercise 2.3.10. Begin with a simple nonrectifiable curve in $\mathbb{R}^2 \subset \mathbb{R}^3$ and build a cone over its image by choosing a point (vertex of the cone) in $\mathbb{R}^3$ and connecting the vertex with every point in the image of the curve by a segment. The intrinsic distance in this cone induced by Euclidean

distance in $\mathbb{R}^3$ is finite for every two points: one can go from one point to the vertex of the cone along a straight segment and then get to the other point along another segment. Show that removing the vertex of the cone makes it disconnected in the topology of induced intrinsic metric, while it is still connected in usual topology.

You can begin with a simple curve whose restriction to any nontrivial interval is nonrectifiable (prove that such a curve does exist). In the original topology this cone is still homeomorphic to a disc. Prove that in the induced intrinsic metric this cone is homeomorphic to the bouquet of a continuum of intervals (that is, the *disjoint* union of segments glued at one point).

Exercise 2.3.11. For two vectors $V, W \in \mathbb{R}^2$, set

$$d(V,W) = \big||V| - |W|\big| + \min(|V|,|W|) \cdot \sqrt{\measuredangle(V,W)},$$

where $\measuredangle(V,W)$ denotes the angle between V and W. Prove that

(i) the topology determined by d is the standard Euclidean one;

(ii) the induced intrinsic metric $\widehat{d}$ is

$$\widehat{d}(V,W) = \begin{cases} \big||V| - |W|\big| & \text{if } \measuredangle(V,W) = 0, \\ \big||V| + |W|\big| & \text{otherwise;} \end{cases}$$

(iii) $(\mathbb{R}^2, \widehat{d})$ is homeomorphic to the bouquet of a continuum of rays.

One can consider the length $L_{\widehat{d}}$ induced by the new metric $\widehat{d}$, and this length in turn determines a "second-stage" intrinsic metric. The reader may imagine how much confusion between various lengths and metrics might arise. However, this is not the case, as the length induced by $\widehat{d}$ is the same as induced by d.

Proposition 2.3.12. *Let (X,d) be a metric space and $\widehat{d}$ be the intrinsic metric induced by d.*

(1) *If γ is a rectifiable curve in (X,d), then $L_{\widehat{d}}(\gamma) = L_d(\gamma)$.*

(2) *The intrinsic metric induced by $\widehat{d}$ coincides with $\widehat{d}$. In other words, inducing a length metric is an idempotent operation.*

Proof. The fact that the length of every curve in (X,d) is not less than the distance between its endpoints implies that $\widehat{d} \geq d$. It follows immediately that $L_{\widehat{d}}(\gamma) \geq L_d(\gamma)$. To prove the inverse inequality, let $[a,b]$ be the domain of γ and let $Y = \{y_i\}$ be an arbitrary partition of $[a,b]$. Observe that $\widehat{d}(\gamma(y_i), \gamma(y_{i+1})) \leq L_d(\gamma, y_i, y_{i+1})$ because the left-hand value is the infimum of lengths one of which is written on the right-hand side. Therefore

$$\Sigma_{\widehat{d}}(Y) = \sum \widehat{d}(\gamma(y_i), \gamma(y_{i+1})) \leq L_d(\gamma).$$

Since Y is an arbitrary partition, the inequality $L_{\widehat{d}}(\gamma) \le L_d(\gamma)$ follows. This proves the first statement of the proposition. The second one is a trivial consequence. □

Remark 2.3.13. The assumption that the curve γ is rectifiable is essential, simply because otherwise it may fail to be continuous in $(X, \widehat{d})$. The set of continuous curves in $(X, \widehat{d})$ is generally a subset of the respective set for (X, d) but it contains all rectifiable curves. See Exercises 2.1.4 and 2.1.5.

2.4. Characterization of Intrinsic Metrics

A metric was said to be intrinsic if it can be obtained by a certain construction. In this chapter we discuss properties that distinguish intrinsic metrics among all metrics and criteria that tell whether a given metric is intrinsic or not.

2.4.1. Another definition of length spaces. The second statement of Proposition 2.3.12 gives a "constructive" criterion to find out whether a given metric can be obtained as an induced intrinsic one. Namely a metric is induced if and only if it induces itself. The following proposition generalizes this for arbitrary intrinsic metrics.

Proposition 2.4.1. *Let (X, d) be a length space and $\widehat{d}$ be the intrinsic metric induced by d. Then $\widehat{d} = d$.*

Proof. Let L be the length function that defines d and L_d be the length induced by d. Observe that $L_d(\gamma) \le L(\gamma)$ for any admissible curve γ of finite length (repeat the respective argument from the proof of Proposition 2.3.12). Obviously a smaller length function determines a smaller metric; thus $\widehat{d} \le d$. On the other hand, we already know that $\widehat{d} \ge d$; hence $\widehat{d} = d$. □

Note that the equality $d = \widehat{d}$ automatically implies that d is an intrinsic metric—just because the induced metric $\widehat{d}$ is always intrinsic. Hence it can be considered as an alternative definition of the term "intrinsic metric". In other words, (X, d) is a length space if and only if for any points $x, y \in X$ and any $\varepsilon > 0$ there exists a curve γ connecting x and y such that $L_d(\gamma) < d(x, y) + \varepsilon$.

If one considers an intrinsic metric d and it does not matter *which* length structure determines it, Proposition 2.4.1 allows us to assume that the length structure is the one induced by d and therefore use all properties of induced length that we established in previous sections.

2.4.2. Recovering a length structure. Now we are ready to answer another natural question, namely: given an intrinsic metric space (X, d), how can we recover the initial length structure L. In fact, recovering the length structure is not possible without additional assumptions because the same intrinsic metric usually can be obtained from many different length structures.

Exercise 2.4.2. Give an example of a length structure on the plane for which all continuous curves are admissible, the resulting intrinsic metric is the standard Euclidean one, but lengths of some curves differ from their Euclidean lengths.

Proposition 2.4.1 shows a natural candidate for the length structure. Namely, the length structure L_d induced by the metric indeed gives rise to the original metric. We will therefore reformulate our question: under what assumptions is $L = L_d$? This kind of question is also important for the following reason: in many examples the initial length function L has many specific features that one may want to exploit. For example, the definition of Finslerian length as the integral of "speed" (see example in Section 2.2) is much more suitable for computing actual lengths than the general definition of induced length applied to the resulting Finslerian metric. If coincidence of two lengths is known, one can combine specific features of the initial function L with general properties of lengths induced by metrics.

Certainly, two length structures may be different just because they have different classes of admissible paths. Thus, we care only for their values on the paths admissible for L. Then we notice that, besides all properties from the definition of length structures, L_d possesses an additional property of lower semi-continuity (see Proposition 2.3.4). Indeed, lower semi-continuity turns out to be the key property here:

Theorem 2.4.3. *If L is a lower semi-continuous length structure, then L coincides with the length structure induced by its intrinsic metric $d = d_L$ on all curves admissible for L: $L(\gamma) = L_d(\gamma)$. As usual, the semi-continuity means that if a sequence of paths $\gamma_i(t)$ pointwise converges to $\gamma(t)$, then* $\liminf L(\gamma_i) \geq L(\gamma)$.

Proof. The inequality $L_d(\gamma) \leq L(\gamma)$ holds for any length structure—see Propositions 2.3.12 and 2.4.1. Let us prove the opposite inequality. By property 2 of length structure, the function $L(t) = L(\gamma_{|_{[a,t]}})$ is uniformly continuous in $[a, b]$ for each rectifiable curve $\gamma : [a, b] \to X$. Hence for every $\varepsilon > 0$, there exists a partition $a = t_0 \leq t_1 \leq \cdots \leq t_{k+1} = b$ such that $d_L(\gamma(t_i), \gamma(t_{i+1})) < \varepsilon$ for every integer i between 0 and k. According to the definition of d_L, for each $i = 0, 1, \ldots, k$ there exists a curve σ_i : $[t_i, t_{i+1}] \to X$ with endpoints $\sigma_i(t_i) = \gamma(t_i)$, $\sigma_i(t_{i+1}) = \gamma(t_{i+1})$ such that

$L(\sigma_i) \leq d(\gamma(t_i), \gamma(t_{i+1})) + \varepsilon/k$. For the concatenation h_ε of the curves σ_i we have

$$L(h_\varepsilon) = \sum_{i=0}^{k} L(\sigma_i) \leq \sum_{i=0}^{k} d_L(\gamma(t_i), \gamma(t_{i+1})) + \varepsilon \leq L_d(\gamma) + \varepsilon.$$

From the triangle inequality one readily sees that $d(\gamma(t), h_\varepsilon(t)) \leq 3\varepsilon$ for every $t \in [a,b]$. It follows that $h_\varepsilon(t) \to \gamma(t)$ since the topology determined by d is finer than the initial one. Now the lower semi-continuity for L implies

$$L(\gamma) \leq \liminf_{\varepsilon \to 0} L(h_\varepsilon) \leq L_d(\gamma),$$

proving the theorem. □

Example 2.4.4. Let us come back to the example from Section 2.2 entitled "Finslerian length". We have chosen a function f in two vector-valued variables and measured the length of a path γ by the formula

$$L(\gamma) = \int_a^b f((\gamma(t)), \gamma'(t))dt.$$

Let us take a function f that does not depend on the first argument and such that $f(x,(1,0)) = f(x,(0,1)) = 1/10$ and $f(x,(1,1)) = 1$. (The fact that f is independent of the first argument intuitively means that the cost of out travel depends on the direction only, and it does not depend on a particular location.) This Finslerian length structure is not lower semi-continuous. To see this, consider a sequence of (stairs-like) broken lines with edges parallel to coordinate axes and approaching the segment $[(0,0),(1,1)]$. The length of this segment (with respect to the Finslerian structure) is $\sqrt{2}$, while the length of each of the broken lines is $1/5$. The reason why this has happen is that we made our travel in the diagonal direction "too inexpensive" compared with the coordinate directions.

Exercise 2.4.5. Show that the Finslerian length structure constructed by a function $f(x,v) = F(v)$ is lower semi-continuous if and only if $F(v)$ satisfies the subadditivity assumption from Definition 1.2.11 of norm: $F(v+w) \leq F(v) + F(w)$. Show that this is also true for general $f(x,v)$: the lower semi-continuity of the corresponding length structure is equivalent to the assumption that f satisfies the inequality $f(x, v+w) \leq f(x,v) + f(x,w)$. This is the reason why, in defining Finslerian structures, one usually requires $f(x,\cdot)$ to be a *norm* for each x.

Exercise 2.4.6. Let d be a Finslerian metric (that is, the metric associated with a Finslerian length structure) on a domain $D \subset \mathbb{R}^n$. Prove that the topology of d coincides with the standard Euclidean one.

2.4.3. Existence of midpoints.

Definition 2.4.7. A point $z \in X$ is called a *midpoint* between points x, y in a metric space (X, d) if $d(x, z) = d(z, y) = \frac{1}{2}d(x, y)$.

The following lemma formulates a necessary condition for a metric to be strictly intrinsic (this condition, under mild assumptions, also turns out to be sufficient).

Lemma 2.4.8. *If d is a strictly intrinsic metric, then for every two points x, y there exists a midpoint z.*

Proof. The length of a shortest path $\gamma : [a, b] \to X$ between x and y is $L(\gamma) = d(x, y)$. Denote $L(t) = L(\gamma_{[a,t]})$. Since $L(t)$ is continuous in t and $L(0) = 0$, there is a $c \in [a, b]$ such that $L(c) = \frac{1}{2}L(b)$. Now choosing $z = \gamma(c)$ and using the fact that the length of a path is not less than the distance between its endpoints, one immediately sees that $d(x, z) = d(y, z) = 1/2d(x, y)$. □

In other words, the previous lemma tells us that if d is strictly intrinsic, then the (closed) balls $\overline{B}_{d(x,y)/2}(x)$ and $\overline{B}_{d(x,y)/2}(y)$ have a nonempty intersection.

Exercise 2.4.9. Show that, for a strictly intrinsic metric d, if $r_1 + r_2 = d(x, y)$, then the balls $\overline{B}_{r_1}(x)$ and $\overline{B}_{r_2}(y)$ have a nonempty intersection.

For intrinsic metrics, an analogous lemma asserts:

Lemma 2.4.10. *If d is an intrinsic metric, then, given a positive ε, for every two points $x, y \in X$ there exists an ε-midpoint z, that is, a point z such that $|2d(x, z) - d(x, y)| \le \varepsilon$ and $|2d(y, z) - d(x, y)| \le \varepsilon$. In other words, if $2r > d(x, y)$, then the balls $B_r(x)$ and $B_r(y)$ have a nonempty intersection.*

Proof. Repeat the same arguments as in the proof of the previous lemma for a path γ connecting x and y, and such that $L(\gamma) - d(x, y) \le \varepsilon$. □

Exercise 2.4.11. Let d be an intrinsic metric. Show that, if $r_1 + r_2 > d(x, y)$, then the balls $B_{r_1}(x)$ and $B_{r_2}(y)$ have a nonempty intersection.

We leave as an exercise the following corollary. In a sense it allows one to measure distances using (sufficiently fine) "dotted lines" between two points instead of measuring distances by means of connecting points by paths (compare also with Definition 2.3.1).

Corollary 2.4.12. *Given a positive ε and two points $x, y \in X$ in a space with a strictly intrinsic metric d, there exists a finite sequence of points $x_1 = x, x_2, \ldots, x_k = y$ such that every two neighboring points in this*

sequence are ε*-close (that is,* $d(x_i, x_{i+1}) \le \varepsilon$ *for all* $I = 1, \dots, k-1$*) and* $\sum_{i=1}^{k-1} d(x_i, x_{i+1}) = d(x, y)$.

For intrinsic metrics the last formula in the corollary should be replaced by $\sum_{i=1}^{k-1} d(x_i, x_{i+1}) - d(x, y) \le \varepsilon$.

Exercise 2.4.13. If x and y are two points in a length space (X, d) and $r < d(x, y)$, then $\operatorname{dist}(y, B_r(x)) = d(x, y) - r$. Prove this.

Exercise 2.4.14. Let X be a length space, Y a metric space, and let a map $f : X \to Y$ be locally Lipschitz with a Lipschitz constant C. Prove that f is Lipschitz with the same constant.

Exercise 2.4.15. Let (X, d) be a length space and A a connected open subset of X. Then d induces on A the (finite-valued) intrinsic metric d_A. Moreover each point $p \in A$ has a neighborhood $U \subset A$ such that for any points p, $q \in U$ we have $d(p, q) = d_A(p, q)$.

2.4.4. Complete intrinsic metrics. In many cases the converse of Lemma 2.4.8 is true: the existence of midpoints (resp. ε-midpoints) implies that a complete metric space is strictly intrinsic (resp. intrinsic). Thus we have a criterion that tells us whether a complete locally-compact metric space is a length space.

Theorem 2.4.16. *Let* (X, d) *be a complete metric space.*

1. If for every $x, y \in X$ *there exists a midpoint, then* d *is strictly intrinsic.*

2. If for every $x, y \in X$ *and every positive* ε *there exists an* ε*-midpoint, then* d *is intrinsic.*

This theorem has the following immediate corollary, which can be used as an alternative criterion for a metric to be intrinsic:

Corollary 2.4.17. *A complete metric space* (X, d) *is a length space iff, given a positive* ε *and two points* $x, y \in X$*, there exists a finite sequence of points* $x_1 = x_1, x_2, \dots, x_k = y$ *such that every two neighboring points in this sequence are* ε*-close (i.e.,* $d(x_i, x_{i+1}) \le \varepsilon$ *for all* $i = 1, \dots, k-1$*) and* $\sum_{i=1}^{k-1} d(x_i, x_{i+1}) < d(x, y) + \varepsilon$.

This corollary says that a metric is intrinsic if and only if, given two points and a positive ε, one can reach one of the points starting from the other one and hopping with jumps shorter than ε and with the total length of the jumps not exceeding the distance between the points plus ε.

Proof of Theorem 2.4.16. To prove that a metric is intrinsic we have to show that for any two points x and y there are paths connecting x and

y whose lengths approximate $d(x, y)$ with any given precision. In case of strictly intrinsic metric, there must be a path whose length is equal to $d(x, y)$. We proceed with the case of strictly intrinsic metrics; modifying this argument for the other case is left as an exercise.

We will construct a path $\gamma : [0, 1] \to X$ between x and y such that $\gamma(0) = x$, $\gamma(1) = y$ and $L(\gamma) = d(x, y)$. First we assign the values of γ for all dyadic rationals (rational numbers of the form $k/2^m$ for some natural numbers k, m). Then we extend this partially defined map by continuity; only this step of the argument will use the completeness of (X, d). Indeed, a path with the desirable properties must pass through a midpoint between x and y. Since such midpoints exist by the assumption of the theorem, choose such a midpoint and assign it to be the image $\gamma(1/2)$. Now we assign $\gamma(1/4)$ to be a midpoint between $x = \gamma(0)$ and $\gamma(1/2)$ and $\gamma(3/4)$ to be a midpoint between $\gamma(1/2)$ and $y = \gamma(1)$. Proceeding this way, we define γ for all dyadic rationals between 0 and 1.

According to our construction, for every two dyadic rationals t_i, t_j

$$d(\gamma(t), \gamma(t')) = |t - t'| \cdot d(x, y). \tag{2.1}$$

This equality implies that the map γ, defined on the set of dyadic rationals, is Lipschitz. Since X is complete and the set of dyadic rationals is dense in $[0, 1]$, this map can be extended to the entire interval $[0, 1]$ (cf. Proposition 1.5.9). Thus we obtained a path $\gamma : [0, 1] \to X$ connecting x and y. Then (2.1) implies that $L(\gamma) = d(x, y)$. □

To see how these results work, we suggest several exercises:

Exercise 2.4.18. Prove that the completion of a length space is a length space.

Exercise 2.4.19. Let X be a compact topological space and let $\{d_n\}_{n=1}^{\infty}$ be a sequence of intrinsic metrics on X that uniformly converge to a metric d (recall that metrics are functions on $X \times X$, so the notion of uniform convergence applies here). Prove that d is intrinsic too.

2.5. Shortest Paths

2.5.1. Curves and natural parameterizations. It is important to remember that, when speaking about paths, we do mean maps and not their images. Indeed, an image of a continuous path may fill a disc; two paths making different number of rounds around a circle are essentially different and in particular have different lengths. Still, changing the parameter by a strictly increasing change of variable means that we visit the same collection of points in the same order; in other words, we traverse the same "curve" in

the same direction. One expects that such changes of variable do not change geometric properties.

This suggests the idea of a curve as a class of equivalent paths with respect to the following relation: paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent if there exists a strictly increasing continuous map φ from I_1 onto I_2 such that $\gamma_1 = \gamma_2 \circ \varphi$. (Check that this indeed is an equivalence relation.) However such a definition would be too restrictive for our purposes. If a path is constant on some subinterval (i.e., it "stops for a while" at a point), so does any path obtained from this one by a change of variable. We want such a path to be equivalent to one that goes the same way except that it passes through the point without stopping. To achieve this, we allow nonstrictly monotone changes of variable. This is formalized by the following

Definition 2.5.1. An (*unparameterized*) *curve* is an equivalence class of the minimal equivalence relation satisfying the following: paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent whenever there exists a nondecreasing continuous map φ from I_1 onto I_2 such that $\gamma_1 = \gamma_2 \circ \varphi$.

Paths (representatives of an equivalence class) are also called *parameterizations* of the curve and *re-parameterizations* of one another.

The term "curve" is used for both unparameterized curves and their parameterizations (i.e., paths). In most cases, "curve" is formally a synonym for "path". However, the former is more appropriate when parameterization-independent properties are considered.

Remark 2.5.2. Existence of a nonstrictly monotone change of variable is not an equivalence relation by itself (due to the lack of inverse changes of variable). This is why we consider the equivalence relation generated by it. In other words, two paths γ and $\bar{\gamma}$ are equivalent (represent the same curve) if and only if there exists a finite sequence of paths $\gamma_1, \gamma_2, \dots, \gamma_n$ such that $\gamma_1 = \gamma$, $\gamma_n = \bar{\gamma}$ and for every $i = 1, \dots, n-1$ either γ_i is obtained from γ_{i+1}, or γ_{i+1} is obtained from γ_i, by a nondecreasing change of variable. This description can be simplified by means of the following exercise.

Exercise 2.5.3. A path $\gamma : I \to X$ is said to be *never-locally-constant* if there exists no interval $[a, b] \subset I$ such that $a \neq b$ and the restriction of γ to $[a, b]$ is a constant map. Prove that

(a) Every curve admits a never-locally-constant parameterization.

Hint: Formally, for a path $\gamma : I \to X$ introduce an equivalence relation $\sim$ on I: $y \sim y'$ if and only if γ is constant on $[y, y']$. Show that the quotient $J = I/\sim$ is homeomorphic to an interval. Then observe that there is a unique map $\tilde{\gamma} : J \to X$ such that $\gamma = \tilde{\gamma} \circ \pi$ where π is the canonical projection I on J, and that $\tilde{\gamma}$ is never-locally-constant and continuous.

(Loosely speaking, J is obtained from I by cutting off all intervals where a path is constant and gluing together the ends of each of the resulting gaps.)

(b) Two paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent if and only if there exist a path $\gamma : J \to X$ and changes of variable $\varphi_1 : I_1 \to J$ and $\varphi_2 : I_2$ $(i = 1, 2)$ such that $\gamma_i = \gamma \circ \varphi_i$.

Hint: Let γ be a never-locally-constant parameterization.

(c) Two paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent if and only if there exist an interval J and changes of variable $\varphi_i : J \to I_i$ $(i = 1, 2)$ such that $\gamma_1 \circ \varphi_1 = \gamma_2 \circ \varphi_2$.

It is easy to see that all parameterizations of a curve have equal lengths (check this). We will denote a curve and its parameterization by the same letter.

Definition 2.5.4. A curve $\gamma : [a, b] \to X$ is called *simple* if the pre-image of every point is an interval.

This means that a simple path is allowed to stop at a point for a while, but having left a point it never comes back. Roughly speaking, the image of a simple path is a curve without "self-intersections". The following very easy exercise justifies the correctness of the definition:

Exercise 2.5.5. If one parameterization of a curve is simple, then so are all other parameterizations.

Exercise 2.5.6. If two simple curves have the same image, then they are equivalent up to a change of variable $t \mapsto -t$.

Our next goal is to choose our favorite parameterization for every curve. These parameterizations are analogous to motion with unit speed in physics or differential geometry:

Definition 2.5.7. A parameterization $\gamma : I \to X$ is *natural* if $L(\gamma, t, t') = t - t'$ for all $t, t' \in I$.

Remark 2.5.8. To verify that a parameterization γ is natural, it suffices to check that $L(\gamma, a, t) = t - a$ for a fixed a and all t. This follows from the formula $L(\gamma, t, t') = L(\gamma, a, t') - L(\gamma, a, t)$.

Other names for natural parameterization are *arc-length parameterization* and *parameterization by arc length*. In other words, a parameterization $\gamma(t)$ is natural when

$$\frac{d}{dt} L(\gamma, a, t) = 1,$$

and one also calls it a *unit speed parameterization*. More generally, one says that a parameterization γ is *of constant speed* v if $L(\gamma, t, t') = v(t - t')$ for all t, t'.

The following proposition tells us that every curve admits a natural parameterization.

Proposition 2.5.9. *Every rectifiable curve $\gamma : [a,b] \to X$ can be represented in the form $\gamma = \bar{\gamma} \circ \varphi$ where $\bar{\gamma} : [0, L(\gamma)] \to X$ is a natural parameterization and φ is a nondecreasing continuous map from $[a,b]$ onto $[0, L(\gamma)]$.*

Proof. The idea of construction of φ is trivial: let $\bar{\gamma}(\tau)$ be the point on (the image of) γ such that the length of the interval of γ between its origin and that point is equal to τ. This is formalized as follows. Define $\varphi(t) = L(\gamma, a, t)$ for all $t \in [a,b]$. Then the function φ is nondecreasing and continuous (by the continuity property of length). The set of its values is the interval $[0, L(\gamma)]$. Now for every $\tau \in [0, L(\gamma)]$ pick a $t \in [a,b]$ such that $\varphi(t) = \tau$ and define $\bar{\gamma}(\tau) = \gamma(t)$. This definition does not depend on the choice of t. Indeed, if $\varphi(t) = \varphi(t')$, then $\gamma(t) = \gamma(t')$ because $L(\gamma, t, t') = \varphi(t') - \varphi(t) = 0$.

Thus we have defined a map $\bar{\gamma} : [0, L(\gamma)] \to X$. The relation $\gamma = \bar{\gamma} \circ \varphi$ follows immediately from the definition. It remains to verify that $\bar{\gamma}$ is continuous and parameterized by arc length. For the former, let $\tau_1 = \varphi(t_1)$ and $\tau_2 = \varphi(t_2)$. Then $\bar{\gamma}(\tau_1)$ and $\bar{\gamma}(\tau_2)$ are the endpoints of the path $\gamma|_{[t_1,t_2]}$. The length of this path is $L(\gamma, t_1, t_2) = \varphi(t_2) - \varphi(t_1) = \tau_2 - \tau_1$. Since the distance between endpoints is no greater than the length, we obtain that $d(\bar{\gamma}(\tau_1), \bar{\gamma}(\tau_2)) \le |\tau_1 - \tau_2|$. This means that $\bar{\gamma}$ is a nonexpanding and hence continuous map. Furthermore $\gamma|_{[t_1,t_2]}$ is a re-parameterization of $\bar{\gamma}|_{[\tau_1,\tau_2]}$ and hence $L(\bar{\gamma}, \tau_1, \tau_2) = L(\gamma, t_1, t_2) = \tau_2 - \tau_1$. Thus $\bar{\gamma}$ is a natural parameterization. □

Exercise 2.5.10. Prove that a natural parameterization of a curve is unique up to a translation $[a,b] \to [a+c, b+c] : t \mapsto t + c$ of the variable.

Exercise 2.5.11. Given $v > 0$, any rectifiable curve admits a parameterization with constant speed v.

The following exercise is a (trivial) corollary of Proposition 2.5.9:

Exercise 2.5.12. If a length space is homeomorphic to a segment, then it is isometric to a segment.

This corollary tells us a remarkable thing: one-dimensional intrinsic geometry is trivial since all intrinsic metrics on a line are locally indistinguishable! We will see that already two-dimensional surfaces are completely different in this respect. By the way, although there is an essentially unique intrinsic metric on $\mathbb{R}$, there are lots of different metrics, which are not intrinsic. For instance, consider $d(x,y) = \sqrt{|x-y|}$ (you may replace the square root by other concave functions to get more examples). The intrinsic distance induced by this metric is infinite everywhere.

2.5.2. Existence of shortest paths. The goal of this section is to prove that a complete locally compact length space is strictly intrinsic, that is, there is a shortest path between every two points.

We begin with the definition of uniform convergence for curves:

Definition 2.5.13. A sequence of curves uniformly converges to a curve γ if they admit parameterizations (with the same domain) that uniformly converge to a parameterization of γ.

The following theorem shows that the space of curves of uniformly bounded lengths in a compact space is compact with respect to the above convergence. This is a version of the Arzela–Ascoli Compactness Theorem in functional analysis.

Theorem 2.5.14 (Arzela–Ascoli Theorem). *In a compact metric space, any sequence of curves with uniformly bounded lengths contains a uniformly converging subsequence.*

Proof. For each γ_i, there is a unique constant speed parameterization by the unit interval $[0,1]$. Uniform boundedness of the lengths of γ_i means that the speeds of these parameterizations are uniformly bounded. In its turn, this implies that for some $C < \infty$

$$d(\gamma_i(t), \gamma_i(t')) \leq L(\gamma, t, t') \leq C|t - t'| \tag{2.2}$$

for every integer I and all $t, t' \in [0,1]$.

Let $S = \{t_j\}$ be a countable dense subset of $[0,1]$. Using the Cantor diagonal process one can find a subsequence γ_{n_i} of $\{\gamma_i\}$ such that for each $j \in \mathbb{N}$ the sequence $\gamma_{n_i}(t_j)$ converges. We plan to show that the subsequence γ_{n_i} itself converges; to avoid double indices, we may assume (without loss of generality) that this subsequence is γ_i itself: for every t_j the limit $\lim_{i\to\infty} \gamma_i(t_j)$ exists.

To prove that the sequence $\{\gamma_i(t)\}$ converges for every $t \in [0,1]$, we will show that this is a Cauchy sequence.

Given $\varepsilon > 0$, choose $t_j \in S$ such that $|t - t_j| < \varepsilon$ and then $N \in \mathbb{N}$ such that $d(\gamma_i(t_j), \gamma_k(t_j)) < \varepsilon$ for all $i, k > N$. For these choices

$$d(\gamma_i(t), \gamma_k(t)) \leq d(\gamma_i(t), \gamma_i(t_j)) + d(\gamma_i(t_j), \gamma_k(t_j)) + d(\gamma_k(t_j), \gamma_k(t)) \leq 3\varepsilon.$$

This proves that $\gamma_i(t)$ is a Cauchy sequence and hence we can define $\gamma(t) = \lim_{j\to\infty} \gamma_j(t_i)$.

Passing to the limit in (2.2) we get

$$d(\gamma(t), \gamma(t')) \leq C|t - t'|, \tag{2.3}$$

and thus γ is a continuous map.

Let us show that γ_i converges to γ uniformly. Given $\varepsilon > 0$, choose $N > \frac{C}{4\varepsilon}$ and let M be such that $d(\gamma(k/N), \gamma_i(k/N)) < \varepsilon/2$ for all $k = 0, 1, \ldots, N$ and all $i > M$. This choice is possible since γ_i converges to γ point-wise. Combining (2.2) with (2.3), for every $0 \le t \le 1$ and $k/N \le t \le (k+1)/N$ we have

$$d(\gamma(t), \gamma_i(k/N)) \le C|t - k/N| + \varepsilon/2 + C|t - k/N| \le \varepsilon$$

for all $i > M$. This concludes the proof. □

Although we already used the notion of shortest paths, this cornerstone notion deserves a formal definition:

Definition 2.5.15. A curve $\gamma : [a, b] \to X$ is a called a *shortest path* if its length is minimal among the curves with the same endpoints, in other words $L(\gamma_1) \ge L(\gamma)$ for any curve γ_1 connecting $\gamma(a)$ and $\gamma(b)$.

It is trivial that any interval of a shortest path is a shortest path (check this).

Remark 2.5.16. In a length space the above definition can be reformulated as follows: a curve $\gamma : [a, b] \to X$ is a shortest path if and only if its length is equal to the distance between its endpoints: $L(\gamma) = d(\gamma(a), \gamma(b))$. Shortest paths in length spaces are also called *distance minimizers.*

Shortest paths in length spaces possess some nice properties that do not hold in general metric spaces. One such property is the following

Proposition 2.5.17. *If shortest paths γ_i in a length space (X, d) converge to a path γ as $i \to \infty$, then γ is also a shortest path.*

Proof. Since the endpoints of γ_i converge to endpoints of γ and the length of each γ_i is equal to the distance between its points, we conclude that $L(\gamma_i) \to d(x, y)$, where x, y are the endpoints of γ. By the lower semi-continuity of length,

$$L(\gamma) \le \lim_{i \to \infty} L(\gamma_i) = d(x, y).$$

□

Exercise 2.5.18. Give an example showing that a limit of shortest paths in a metric space may fail to be a shortest path.

The example of $\mathbb{R}^2 \setminus \{0\}$ shows that there may be no shortest path between two points. On the other hand, there may be several different shortest paths between the same two points: for instance, consider two antipodal points in a sphere.

Convention. We use the notation $[x, y]$ to denote a shortest path between points x and y. This notation is convenient when either such shortest path is unique or it does not matter which of the shortest paths is considered. This notation is well compatible with our notation for segments in Euclidean space, since the latter are just shortest paths with respect to Euclidean distance.

Proposition 2.5.19. *Let (X, d) be a compact metric space and let $x, y \in X$ be points that can be connected by at least one rectifiable curve. Then there exists a shortest path between x and y.*

Proof. Let L_{inf} denote the infimum of lengths of rectifiable curves connecting x and y. Then there exists a sequence $\{\gamma_i\}$ of such curves with $L(\gamma_i) \to L_{\text{inf}}$. According to Theorem 2.5.14, the sequence $\{\gamma_i\}$ contains a converging subsequence. Without loss of generality we may assume that $\{\gamma_i\}$ itself converges to a curve γ. Then γ has the same endpoints and, by the lower semi-continuity of length, $L(\gamma) \le \lim L(\gamma_i) = L_{\text{inf}}$. Thus $L(\gamma) = L_{\text{inf}}$. □

Corollary 2.5.20. *Let (X, d) be a boundedly compact metric space. Then for every two points $x, y \in X$ connected by a rectifiable curve there exists a shortest path between x and y.*

Proof. Let L be a length of some rectifiable curve connecting x and y. Observe that this curve, as well as any shorter curve with the same endpoints, is contained in the closed metric ball of radius L centered at x. So it is sufficient to prove the existence of a shortest path only inside this ball. Since the ball is compact, this follows from the previous proposition. □

Definition 2.5.21. A topological space X is called *locally compact* if every point of X has a pre-compact neighborhood.

Proposition 2.5.22. *If (X, d) is a complete locally compact length space, then every closed ball in X is compact (and hence X is boundedly compact).*

Note that in this proposition it is essential that X is a length space. For example, a space where all distances between points equal to 1 is locally compact and complete, but a closed unit ball in such a space is not compact unless the space has a finite cardinality.

Proof of Proposition 2.5.22. Let $x \in X$ be an arbitrary point. Observe that if $\bar{B}_r(x)$ is compact for some r, then $\bar{B}_\rho(x)$ is compact for any $\rho < r$. Define

$$R = \sup\{r > 0 : \bar{B}_r(x) \text{ is compact}\}.$$

Since x has a pre-compact neighborhood, we have $R > 0$. Suppose that $R < \infty$ and denote the ball $\bar{B}_R(x)$ by B.

First let us prove that B is compact. Since B is a closed set in a complete space, it suffices to prove that for any $\varepsilon > 0$ it contains a finite ε-net. We may assume that $\varepsilon < R$. Let B' denote the ball $\bar{B}_{R-\varepsilon/3}(x)$. This ball is compact and hence it contains a finite $(\varepsilon/3)$-net S. Let $y \in B$. Since X is a length space, we have $\operatorname{dist}(y, B') \le \varepsilon/3$. Therefore there exists a point $y' \in B'$ with $d(y, y') < \varepsilon/2$. On the other hand, $\operatorname{dist}(y', S) \le \varepsilon/2$; hence $\operatorname{dist}(y, S) < \varepsilon$. This means that S is an ε-net for B, and we have proven the compactness of B.

Every point $y \in B$ has a pre-compact neighborhood U_y. Pick a finite collection $\{U_y\}_{y \in Y}$ of such neighborhoods that cover B. Their union $U = \bigcup_{y \in Y} U_y$ is a pre-compact neighborhood of B. Using the compactness of B again, we can conclude that there exists a positive $\varepsilon > 0$ such that the ε-neighborhood of B is contained in U. Since X is a length space, the ε-neighborhood of B is the ball $B_{R+\varepsilon}(x)$, and its closure is $\bar{B}_{R+\varepsilon}(x)$. Therefore $\bar{B}_{R+\varepsilon}(x)$ is compact. This contradicts the choice of R.

Thus the assumption $R < \infty$ is wrong; hence $R = \infty$. This means that all balls centered at x are compact. □

Combining Proposition 2.5.22 and the previous Corollary 2.5.20, we obtain the main result of this section.

Theorem 2.5.23. *Let (X, d) be a complete locally compact length space. Then this space is strictly intrinsic: for every $x, y \in X$ such that $d(x, y) < \infty$ there exists a shortest path γ connecting x and y, i.e., a curve $\gamma : [a, b] \to X$ such that $\gamma(a) = x$, $\gamma(b) = y$ and $L(\gamma) = d(x, y)$.*

The example of $\mathbb{R}^2 \setminus \{0\}$ shows that completeness in this theorem is essential. So is local compactness:

Exercise 2.5.24. Give an example of a complete length space (with finite metric) in which there is no shortest path between some points.

However, the condition that X is a length space in the above considerations can be omitted, at the expense of more complicated formulations:

Exercise 2.5.25. Let X be a complete locally compact metric space (not necessarily a length space). Prove that for every two points $x, y \in X$ that can be connected by a rectifiable curve, there exists a shortest path between x and y.

Hint: Prove that for every $R > 0$ the set of points that can be connected to x by a curve of length less than R is pre-compact. In other words, balls of induced length metric are pre-compact in the topology of the original metric (!). To prove this, modify the proof of Proposition 2.5.22.

Exercise 2.5.26. Is it true that a completion of a locally compact length space is locally compact?

2.5.3. Geodesics and the Hopf–Rinow Theorem.

Definition 2.5.27. Let X be a length space. A curve $\gamma: I \to X$ is called a *geodesic* if for every $t \in I$ there exists an interval J containing a neighborhood of t in I such that $\gamma|_J$ is a shortest path. In other words, a geodesic is a curve which is locally a distance minimizer (i.e., a shortest path).

The example of the sphere shows that there are geodesics that are not shortest paths: whereas every segment of a great circle on a sphere is a geodesic, it fails to be shortest as soon as it is longer than half of the equator. Although the spherical example suggests that a shortest path between two points should be unique at least locally, this is also not true in general. To see an example, consider the surface of a cube with its intrinsic metric induced by the Euclidean metric of its ambient space. We suggest that the reader shows that any neighborhood of a vertex contains points with multiple shortest paths between them. Another natural conjecture that turns out to be wrong is that a limit of geodesics is a geodesic; give a counterexample to this on the surface of the cube. We will see later that these phenomena are caused by nonsmoothness of the surface and that indeed such things never happen to smooth examples.

It is clear that a shortest path always admits a natural parameterization, and hence so does a geodesic.

Intuitively, a space is noncomplete if a point is missing. One may suspect that this absence of a point can be noticed by moving along a geodesic that would end at this point: the interval for which our motion is well-defined is not closed. The theorem below formalizes this observation.

Before formulating it, we generalize the notion of a shortest path so as to include paths defined on nonclosed intervals. Namely we say that a curve $\gamma: I \to X$ (where $I \subset \mathbb{R}$ is an interval and X is a metric space) is a shortest path, or a *minimal geodesic*, if its restriction to any interval $[a, b] \subset I$ is a shortest path in the sense of Definition 2.5.15.

Theorem 2.5.28 (Hopf–Rinow–Cohn-Vossen Theorem). *For a locally compact length space X, the following four assertions are equivalent:*

(i) *X is complete.*

(ii) *X is boundedly compact, i.e., every closed metric ball in X is compact.*

(iii) *Every geodesic $\gamma: [0, a) \to X$ can be extended to a continuous path $\overline{\gamma}: [0, a] \to X$.*

(iv) *There is a point* $p \in X$ *such that every shortest path* $\gamma\colon [0,a) \to X$ *with* $\gamma(0) = p$ *can be extended to a continuous path* $\overline{\gamma}\colon [0,a] \to X$.

The theorem generalizes the classical Hopf-Rinow theorem which originally was proved only in smooth situation, i.e., for Riemannian manifolds.

Remark 2.5.29. By Theorem 2.5.23, these conditions imply that every two points a, b can be connected by a shortest path.

Proof of the theorem. Implications (ii) $\Longrightarrow$ (i) $\Longrightarrow$ (iii) $\Longrightarrow$ (iv) are left as easy exercises. We will prove that (iv) implies (ii). The proof uses the same general scheme as the one of Proposition 2.5.22. The details are slightly more complicated because we cannot utilize completeness at this point. We can use only the property (iv) and this requires a more delicate argument.

Since X is locally compact, sufficiently small closed balls $\overline{B}_r(p)$ are compact. Reasoning by contradiction (that is, assuming that there are noncompact closed balls), define

$$R = \sup\{r : \overline{B}_r(p) \text{ is a compact set }\}$$

and assume that $R < \infty$. The argument consists of two steps.

1. First, we prove that the open ball $B_R(p)$ is pre-compact. To do this, it suffices to show that every sequence $\{x_i\}$ in this ball contains a converging subsequence (whose limit does not necessarily belong to this ball). Set $r_i = d(p, x_i)$. One may assume that $r_i \to R$ as $i \to \infty$; otherwise a subsequence of $\{x_i\}$ is contained in a ball $\overline{B}_r(p)$ for some $r < R$, and then there is a converging subsequence because this smaller ball is compact (by the choice of R).

Let $\gamma_i\colon [0, r_i] \to X$ be a (naturally parameterized) shortest path connection p to x_i. Such a shortest path exists because x_i belongs to a compact ball centered at p (see the proof of Corollary 2.5.20). We can choose a subsequence of $\{\gamma_i\}$ such that the restrictions of the paths to $[0, r_1]$ converge (along this subsequence). From this subsequence, we choose a further subsequence of paths whose restrictions to $[0, r_2]$ converge, and so on. Then the Cantor diagonal procedure (that is, picking the nth element from the nth subsequence for $n = 1, 2, \dots$) yields a sequence $\{\gamma_{i_n}\}$ such that for every $t \in [0, R)$ the sequence $\{\gamma_{i_n}(t)\}$ converges in X. (More precisely, points $\gamma_{i_n}(t)$ are well-defined for all large enough n and form a converging sequence.)

Define $\gamma(t) = \lim \gamma_{i_n}(t)$; then $\gamma\colon [0, R) \to X$ is a nonexpanding map and a shortest path (see Proposition 2.5.17). By the assertion (iv), there is a continuous extension $\overline{\gamma}\colon [0, R] \to X$. One easily sees (exercise!) that the points x_{i_n} (i.e., the endpoints of our converging curves $\{\gamma_{i_n}\}$) converge to $\overline{\gamma}(R)$.

2. Since the open ball $B_R(p)$ is pre-compact, the closed ball $\overline{B}_R(p)$ is compact. (Recall that a closed ball in a length space is the closure of the respective open ball.) Now we show that a ball $B_{R+\varepsilon}(p)$ is pre-compact for some $\varepsilon > 0$. Since X is locally compact, for every $x \in \overline{B}_R(p)$ there is an $r(x) > 0$ such that the ball $B_{r(x)}(x)$ is pre-compact. Choose a finite subcover $\overline{B}_{r(x_i)}(x_i)$ out of the cover of $\overline{B}_R(x)$ by these balls. The union of these balls is pre-compact and contains the ball $B_{R+\varepsilon}(p)$ for $\varepsilon = \min r_i > 0$. This contradicts the choice of R. □

2.6. Length and Hausdorff Measure

Recall that the length of a curve is independent of the parameterization. This fact suggests that the length of a simple curve can be recovered from its image in the space, i.e., the set of points it passes through. In this section we show that the length of a path actually equals the one-dimensional Hausdorff measure of the image. We assume that the normalization constant $C(1)$ for Hausdorff measure is 1 (for a definition and elementary properties of Hausdorff measure see Section 1.7).

Lemma 2.6.1. *If X is a connected metric space, then $\mu_1(X) \geq \operatorname{diam} X$.*

Proof. 1. A general observation: in the definition of Hausdorff measure we can restrict ourselves to coverings by *open* sets S_i. Indeed, an arbitrary covering $\{S_i\}$ can be replaced by an open covering $\{S'_i\}$ where

$$S'_i = U_{\delta/2^i}(S_i) := \{x \in X : \operatorname{dist}(x, S_i) < \delta/2^i\}$$

for a small $\delta > 0$. Then $\operatorname{diam} S'_i \leq \operatorname{diam} S_i + 2\delta/2^i$ and hence $w_1(\{S'_i\}) \leq w_1(\{S_i\}) + 2\delta$. Since δ is arbitrary, it follows that the measure can be approximated by 1-weights of open coverings. (*Exercise*: extend the argument to work for measures of all dimensions.)

2. Let X be a connected topological space and $\{S_i\}$ an open covering of X. Then for every two points $x, y \in X$ there exists a finite sequence $S_{i_1}, \dots, S_{i_n}$ of different sets such that $x \in S_{i_1}$, $y \in S_{i_n}$, and $S_{i_k} \cap S_{i_{k+1}} \neq \emptyset$ for all k, $1 \leq k \leq n-1$. To prove this, fix a point $x \in X$ and consider the set Y of all points $y \in X$ for which such a sequence exists. It is clear that for every set S_i one has either $S_i \subset Y$ or $S_i \subset X \setminus Y$. Therefore both Y and $X \setminus Y$ are open sets. Since X is connected, it follows that $Y = X$ and therefore any point $y \in Y$ is "accessible" by a sequence of sets as described above.

3. Let $\{S_i\}$ be an open covering of X, x and y two points of X, and $S_{i_1}, \dots, S_{i_n}$ a sequence from Step 2. For $k = 1, \dots, n-1$ pick a point $x_k \in S_{i_k} \cap S_{i_{k+1}}$, and define $x_0 = x$, $x_n = y$. Then $|x_{k-1}x_k| \leq \operatorname{diam} S_{i_k}$ for

all $k = 1, \dots, n$ because both x_{k-1} and x_k are in S_{i_k}. Hence

$$\sum \operatorname{diam} S_i \ge \sum_{k=1}^{n} \operatorname{diam} S_{i_k} \ge \sum_{k=1}^{n} |x_{k-1}x_k| \ge |x_0x_n| = |xy|.$$

Due to the observation made in step 1, it follows that $\mu_1(X) \ge |xy|$. Since x and y are arbitrary points, this means that $\mu_1(X) \ge \operatorname{diam} X$. □

Theorem 2.6.2. *Let X be a metric space, $\gamma\ [a,b] \to X$ a simple curve. Then $L(\gamma) = \mu_1(\gamma([a,b]))$.*

Proof. Let $S = \gamma([a,b])$ and $L = L(\gamma)$. Without loss of generality assume that γ is parameterized by arc length: $\gamma : [0,L] \to X$. Then for every natural number N, S is covered by N intervals of γ each of length L/N, namely by the sets $\gamma([i\frac{L}{N}, (i+1)\frac{L}{N}])$, $i = 0, 1, \dots, N-1$. The diameters of these sets are no greater than L/N. Hence the sum of the diameters is no greater than L, whereas the diameters themselves approach 0 as $N \to \infty$. This shows that $\mu_1(S) \le L(\gamma)$.

On the other hand, for a partition $a = t_0 \le t_1 \le \dots \le t_n = b$ of $[a,b]$, let $S_i = \gamma([t_i, t_{i+1}])$ for $i = 1, \dots, n-1$. The sets S_i are disjoint modulo a finite number of points $\gamma(t_i)$. Since the one-dimensional measure of a single point is obviously zero, one has $\mu_1(S) = \sum \mu_1(S_i)$. By Lemma 2.6.1, $\mu_1(S_i) \ge \operatorname{diam} S_i \ge |\gamma(t_i)\gamma(t_{i+1})|$, and hence

$$\mu_1(S) = \sum \mu_1(S_i) \ge \sum |\gamma(t_i)\gamma(t_{i+1})|$$

for any partition $\{t_i\}$. This implies that $\mu_1(S) \ge L(\gamma)$. □

Remark 2.6.3. If γ is not simple, the same argument shows that $L(\gamma) \ge \mu_1(\gamma([a,b]))$.

Exercise 2.6.4. Prove that, for any curve $\gamma : [a,b] \to X$,

$$\ell(\gamma) = \sum_{k \in \mathbb{N} \cup \{\infty\}} k \cdot \mu_1(\{x \in X : \#(\gamma^{-1}(x)) = k\})$$

where # denotes the cardinality and $\infty \cdot 0 = 0$.

Hint: The right-hand part is the integral of the function $x \mapsto \#(\gamma^{-1}(x))$ with respect to the measure μ_1. Combine Levy's limit theorem for integrals with inequalities "$\operatorname{diam} \le \mu_1 \le \ell$" applied to small pieces of γ.

2.7. Length and Lipschitz Speed

This section is rather technical. We will generalize the formula "length equals the integral of speed" which is well known and trivial for smooth curves in $\mathbb{R}^n$ (cf. Exercise 2.3.3). The last theorem of this section requires the knowledge of Lebesgue integration, and we will use without proofs some facts from measure theory.

Definition 2.7.1. Let (X, d) be a metric space and $\gamma : I \to X$ a curve. The *speed* of γ at $t \in I$, denoted by $v_\gamma(t)$, is defined by

$$v_\gamma(t) := \lim_{\varepsilon \to 0} \frac{d(\gamma(t), \gamma(t+\varepsilon))}{|\varepsilon|}$$

if the limit exists.

Exercise 2.7.2. Let γ be a differentiable curve in $\mathbb{R}^n$ (more generally, in a normed vector space). Prove that $v_\gamma(t)$ exists for all t and $v_\gamma(t) = |\gamma'(t)|$.

Exercise 2.7.3. Let X be a metric space and $\gamma : [a, b] \to X$ a curve. Suppose that $v_\gamma(t)$ exists for all $t \in [a, b]$ and is continuous in t. Prove that

$$L(\gamma) = \int_a^b v_\gamma(t)\, dt$$

(compare Exercise 2.3.3).

Of course, the speed of a curve may not exist. However, it exists almost everywhere (i.e., except a set of zero measure) for a wide class of curves, namely for every Lipschitz curve. (Note that every rectifiable curve admits Lipschitz parameterization. For example, all natural parameterizations are Lipschitz with unit Lipschitz constant.) Moreover, the length of a Lipschitz curve equals the (Lebesgue) integral of its speed.

As a first step, we prove the following

Theorem 2.7.4. *Let (X, d) be a metric space, $\gamma : [a, b] \to X$ a rectifiable curve. Then for almost all $t \in [a, b]$ (i.e., for all t except a set of zero measure) the following holds: either*

$$\liminf_{\varepsilon, \varepsilon' \to 0+} \frac{L(\gamma|_{[t-\varepsilon, t+\varepsilon']})}{\varepsilon + \varepsilon'} = 0,$$

or

$$\lim_{\varepsilon, \varepsilon' \to 0+} \frac{d(\gamma(t-\varepsilon), \gamma(t+\varepsilon'))}{L(\gamma|_{[t-\varepsilon, t+\varepsilon']})} = 1.$$

One can let ε or ε' in the above formulas be zero. This gives the following

Corollary 2.7.5. *If γ is as in Theorem 2.7.4, then for almost all $t \in [a, b]$ either*

$$\liminf_{\varepsilon \to 0} \frac{L(\gamma|_{[t, t+\varepsilon]})}{|\varepsilon|} = 0,$$

or

$$\lim_{\varepsilon \to 0} \frac{d(\gamma(t), \gamma(t+\varepsilon))}{L(\gamma|_{[t, t+\varepsilon]})} = 1$$

(if $\varepsilon < 0$, the interval $[t, t+\varepsilon]$ in the denominator of the last formula should be interpreted as $[t+\varepsilon, t]$).

Proof of Theorem 2.7.4. Suppose the contrary. For every $\alpha > 0$ let Z_α denote the set of all $t \in [a,b]$ such that

$$\liminf_{\varepsilon,\varepsilon'\to 0+} \frac{L(\gamma|_{[t-\varepsilon,t+\varepsilon']})}{\varepsilon+\varepsilon'} > \alpha$$

and

$$\liminf_{\varepsilon,\varepsilon'\to 0+} \frac{d(\gamma(t-\varepsilon),\gamma(t+\varepsilon'))}{L(\gamma|_{[t-\varepsilon,t+\varepsilon']})} < 1-\alpha.$$

Then $\mu_1(Z_\alpha) > 0$ for all sufficiently small α. Indeed, otherwise the set $Z_0 = \bigcup_{\alpha>0} Z_\alpha = \bigcup_{n\in\mathbb{N}} Z_{1/n}$ would have zero measure, and this is equivalent to the statement of the theorem. Fix an $\alpha > 0$ such that $\mu_1(Z_\alpha) > 0$. For brevity, we denote $Z = Z_\alpha$ and $\mu = \mu_1(Z)$. Choose ε_0 so small that for any partition $\{y_i\}_{i=1}^N$ ($a = y_0 \le y_1 \le \dots \le y_N = b$) of $[a,b]$ such that $\max_i(y_i - y_{i-1}) < \varepsilon_0$, one has

$$L(\gamma) - \sum_{i=1}^N d(\gamma(y_{i-1}),\gamma(y_i)) < \mu\alpha^2/2.$$

Such an ε_0 exists by Exercise 2.3.2. Consider the set $\mathfrak{B}$ of all intervals of the form $[t-\varepsilon, t+\varepsilon']$ such that $t \in Z$, $\varepsilon+\varepsilon' < \varepsilon_0$, $L(\gamma|_{[t-\varepsilon,t+\varepsilon']}) > \alpha(\varepsilon+\varepsilon')$ and

$$d(\gamma(t-\varepsilon),\gamma(t+\varepsilon')) < (1-\alpha)L(\gamma|_{[t-\varepsilon,t+\varepsilon']}).$$

By the definition of $Z = Z_\alpha$, every point $t \in Z$ is contained in arbitrarily short elements of $\mathfrak{B}$. Applying Vitali's covering theorem (Theorem 1.7.14) we can extract from $\mathfrak{B}$ a countable collection $\{[t_i-\varepsilon_i, t_i+\varepsilon_i']\}_{i=1}^\infty$ of disjoint intervals that covers Z up to a set of zero measure. In particular,

$$\sum_{i=1}^\infty(\varepsilon_i+\varepsilon_i') = \mu_1\Big(\bigcup[t_i-\varepsilon_i,t_i+\varepsilon_i']\Big) \ge \mu_1(Z) = \mu.$$

Hence for a sufficiently large M,

$$\sum_{i=1}^M(\varepsilon_i+\varepsilon_i') > \mu/2.$$

Since the intervals $\{[t_i-\varepsilon_i, t_i+\varepsilon_i']\}_{i=1}^M$ are disjoint, they can be included in a partition $\{y_j\}_{j=1}^N$ all whose intervals are shorter than ε_0. We denote $L_j = L(\gamma|_{[y_{j-1},y_j]})$ and $d_j = d(\gamma(y_{j-1}),\gamma(y_j))$. By the choice of ε_0, we have

$$\sum_{j=1}^N(L_j-d_j) = L(\gamma) - \sum_{j=1}^N d_j < \mu\alpha^2/2.$$

In the left-hand sum above all terms are nonnegative and those for which $[y_{j-1},y_j] \in \mathfrak{B}$ (i.e., $y_{j-1} = t_i-\varepsilon_i$ and $y_j = t_i+\varepsilon_i'$ for some i) satisfy

$$L_j - d_j > \alpha L_j > \alpha^2(y_{j-1}-y_j) = \alpha^2(\varepsilon_i+\varepsilon_i').$$

Therefore
$$\sum_{j=1}^{N}(L_j - d_j) \geq \alpha^2 \sum_{i=1}^{M}(\varepsilon_i + \varepsilon_i') > \mu\alpha^2/2.$$
This contradiction proves the theorem. □

Theorem 2.7.6. *Let X be a metric space; $\gamma : [a,b] \to X$ is a Lipschitz curve. Then the speed $v_\gamma(t)$ exists for almost all $t \in [a,b]$ and $L(\gamma) = \int_a^b v_\gamma(t)\,dt$ where $\int$ is the Lebesgue integral.*

Proof. We need the following fact ([**Fe**], Theorem 2.9.19): if $f : [a,b] \to \mathbb{R}$ is a Lipschitz function, then the derivative $f'(t)$ exists for almost all $t \in [a,b]$ and $\int_a^b f'(t)\,dt = f(b) - f(a)$. (*Remark*: the proof of this fact is based on the same ideas as the above proof of Theorem 2.7.4 though it is more complicated.)

Define $f(t) = L(\gamma|_{[a,t]})$ for $t \in [a,b]$. Then f is a Lipschitz function and hence is differentiable almost everywhere. We rewrite $f'(t)$ as follows:
$$f'(t) = \lim_{\varepsilon\to 0} \frac{L(\gamma|_{[t,t+\varepsilon]})}{|\varepsilon|} = \lim_{\varepsilon\to 0} \frac{L(\gamma|_{[t,t+\varepsilon]})}{d(\gamma(t),\gamma(t+\varepsilon))} \cdot \frac{d(\gamma(t),\gamma(t+\varepsilon))}{|\varepsilon|}.$$
By Corollary 2.7.5, for almost all $t \in [a,b]$ either $f'(t) = 0$ or the first term in the last product goes to 1 as $\varepsilon \to 0$. In the first case we have
$$v_\gamma(t) = \lim_{\varepsilon\to 0} \frac{d(\gamma(t),\gamma(t+\varepsilon))}{|\varepsilon|} = 0$$
because $d(\gamma(t),\gamma(t+\varepsilon)) \leq L(\gamma|_{[t,t+\varepsilon]})$. In the second case, it follows that
$$v_\gamma(t) = \lim_{\varepsilon\to 0+} \frac{d(\gamma(t),\gamma(t+\varepsilon))}{|\varepsilon|} = f'(t).$$
Thus $v_\gamma(t)$ exists and equals $f'(t)$ in both cases. The theorem follows. □

Exercise 2.7.7. Give an example of a nonconstant curve γ in $\mathbb{R}^2$ for which $v_\gamma = 0$ almost everywhere.

Exercise 2.7.8. Let X be a metric space and $\gamma : I \to X$ a curve. For a $t \in [a,b]$ define
$$\operatorname{dil}_t(\gamma) = \limsup_{\varepsilon\to 0+} \operatorname{dil}(\gamma|_{[t-\varepsilon,t+\varepsilon]}).$$

(1) Prove that $\operatorname{dil}_t(\gamma) \geq v_\gamma(t)$ whenever $v_\gamma(t)$ is defined.

(2) Prove that, if $v_\gamma(t)$ is defined for all t and is continuous in t, then $\operatorname{dil}_t(\gamma) = v_\gamma(t)$ for all t.

(3) Give an example where γ is Lipschitz but $\int_a^b \operatorname{dil}_t(\gamma) \neq L(\gamma)$.

Chapter 3

Constructions

3.1. Locality, Gluing and Maximal Metrics

3.1.1. Locality. First we observe that to reconstruct an intrinsic metric it is enough to know it only locally, as the following proposition asserts:

Lemma 3.1.1. *Let a topological space X be covered by a collection (not necessarily finite or countable) of open sets $\{X_\alpha\}$. Assume that each X_α is equipped with a length structure L_α and that these structures agree. The latter assumption means that, if a curve γ belongs to the intersection of X_α and X_β, then $L_\alpha(\gamma) = L_\beta(\gamma)$.*

Then there exists a unique length structure L on X whose restriction to every X_α is L_α. Moreover, if X is connected and all intrinsic metrics induced by L_α on X_α are finite, then so is L.

Proof. Consider a curve $\gamma : [a, b] \to X$. The inverse images $\gamma^{-1}(X_\alpha)$ form an open covering of $[a, b]$. The compactness of $[a, b]$ implies that there is a finite partition $a = t_0 \le t_1 \le \cdots \le t_n = b$ such that every segment $[t_i, t_{i+1}]$ is contained in an element of this covering. Then every image $\gamma([t_i, t_{i+1}])$ is contained in one of the sets X_α, and the length of $\gamma_{|_{[t_i,t_{i+1}]}}$ is given by L_α. By the additivity of length, the length of γ must be equal to the sum of lengths of its restrictions to the intervals $[t_i, t_{i+1}]$. This proves the uniqueness part of the lemma and suggests how to define L. To complete the proof one checks that $L(\gamma)$ defined this way is independent of the choice of a partition and satisfies the conditions from the definition of length structures. This part is left to the reader as an exercise.

To prove the statement about finiteness, fix a point $x \in X$ and consider the set Y of all points whose distances from x are finite. Every set X_α is

contained in either Y or $X \setminus Y$; hence both Y and $X \setminus Y$ are open. Since X is connected, it follows that $Y = X$ and the metric is finite. □

Intrinsic metrics also enjoy the following property of *locality*:

Corollary 3.1.2. *Consider two intrinsic metrics d_1 and d_2 defined on the same set X and inducing the same topology. Assume that every point $x \in X$ has a neighborhood U_x such that the restrictions of the metrics to this neighborhood coincide: for every $p, q \in U_x$, $d_1(p,q) = d_2(p,q)$. Then $d_1 = d_2$.*

The corollary implies that, unlike general metrics, an intrinsic metric can be recovered from local measurements.

Exercise 3.1.3. Prove the corollary.

Exercise 3.1.4. Give an example demonstrating that Corollary 3.1.2 fails without the assumption that the metrics in question are intrinsic.

The next proposition shows that locality in fact characterizes intrinsic metrics among all metrics:

Proposition 3.1.5. *If a complete metric d on a set X is not intrinsic, then there exists another metric d_1 on X such that $d \neq d_1$ but every point has a neighborhood where d and d_1 coincide.*

Proof. For every positive ε, introduce the metric d_ε:

$$d_\varepsilon(x,y) = \inf \sum_{i=0}^{k} d(p_i, p_{i+1}),$$

where the infimum is taken over all finite sequences of points $p_0, p_1, \dots, p_{k+1}$ such that $p_0 = x$, $p_{k+1} = y$ and $d(p_i, p_{i+1}) \le \varepsilon$ for all $i = 0, 1, \dots, k$. Intuitively, d_ε is measured by leaping from x to y with jumps each no longer than ε. Obviously $d_\varepsilon(x,y) = d(x,y)$ if $d(x,y) \le \varepsilon$, and thus d and d_ε coincide on every ball of radius $\varepsilon/2$. On the other hand, if $d_\varepsilon = d$ for all $\varepsilon > 0$, then d is intrinsic by Proposition 2.4.17. □

3.1.2. Gluing. Imagine that we glue together two length spaces, or even some points in the same length space, as we do to glue a Möbius strip or a ring out of a paper rectangle. How do we measure distances in this space? For two points "on different sides of the gluing line", the first thing that occurs to one's mind would be to look for a shortest path that consists of two pieces: first it goes in one space to the gluing edge and then from this edge it continues in the other space. In other words, one thinks of a path with a "gap", but the gap is eliminated by gluing. As the matter of fact,

this is not at all so simple. It may very well happen that a shortest path has to cross the gluing edge several or even infinitely many times.

Before passing to rigorous definitions, let us consider several examples.

Example 3.1.6. Begin with a strip $\mathbb{R} \times [0,1] \subset \mathbb{R}^2$ and glue together its edges by identifying $(x, 1)$ with $(x+100, 0)$ for all x. We obtain a topological cylinder.

Question. Is the distance between $(0, 1/2)$ and $(1000, 1/2)$ (after the identifications are made) greater than 899?

Answer. No, it is even less than 11! Indeed, consider the path whose itinerary is

$|(0,1/2) \to (0,1)| = 1/2$, $|(0,1) \to (100,0)| = 0$ ("free"),

$|(100,0) \to (100,1)| = 1$, $|(100,1) \to (200,0)| = 0$ ("free"),

$\dots,$

$|(900,0) \to (900,1)| = 1$, $|(900,1) \to (1000,0)| = 0$ ("free"),

$|(1000,0) \to (1000,1/2)| = 1/2$.

The length of this path is 10.

Exercise 3.1.7. Is this a shortest path?

Exercise 3.1.8. Consider the region $E = \{(x,y)\colon 0 \le y \le e^{-x}, x \ge 0\}$. (This is the region enclosed between the graph of e^{-x}, $x \ge 0$, and the x-axis.) We glue together the infinite edges of this region by identifying $(x, 0)$ with $(x+1, e^{x+1})$ for all $x \ge 0$. Show that the diameter of the resulting space is finite!

Hint: Consider the path consisting of the segments $\{[(n,0),(n,e^{-n})],$ $n = 1, 2, \dots\}$. Observe that this is indeed a continuous path and its length is equal to

$$\sum_{m=1}^{\infty} e^{-m} < \int_0^{\infty} e^{-x} dx = 1.$$

Remark 3.1.9. This is not an artificial example: such examples naturally arise in hyperbolic geometry.

Certainly, both the arguments and the questions themselves are informal: we have not given a definition of gluing yet. To show that this notion is not so simple and indeed needs a rigorous definition, look at the following more striking example:

Example 3.1.10. Begin with $\mathbb{R}^2$ and glue each point (x, y) with $(-y, 2x)$.

Question 1. If we forget about metric and do this identification topologically, what is the quotient space?

Question 2. How can we find the distance between two points after this identification?

Answer. Very simple: the distance is identically zero.

Exercise 3.1.11. Prove that, given a positive ε and two arbitrary points in the space from the above example, there is a path between the points whose length is less than ε.

Hint: Our gluing rule can be re-written as

$$(x,y) \to \left(\frac{y}{2}, -x\right) \to \left(\frac{-x}{2}, \frac{-y}{2}\right) \to \left(\frac{-y}{4}, \frac{x}{2}\right) \to \left(\frac{x}{4}, \frac{-y}{4}\right) \to \dots .$$

These examples suggest the following strategy of measuring the metric resulting from "gluing some points together" in terms of the intact space (before identifications are made). We consider finite sequences of paths such that the terminating point of each of them will be glued with the starting point of the next one. To measure the distance between two points we require that the first path starts at one of the points and the last path ends at the other. The infimum of lengths of such paths gives us the new distance. In other words, to measure the distance between x and y, we connect x and y by a finite sequence of points (a "dotted line"). If we added the distances between neighboring points in this sequence and took the infimum of such sums, this would be just the original distance between x and y. What we do instead is we add only the distances between neighboring points that are nonequivalent. Intuitively, we may jump from a point to an equivalent point for free.

Definition 3.1.12. Let (X,d) be a metric space and let R be an equivalence relation on X. The *quotient semi-metric* d_R is defined as

$$d_R(x,y) = \inf\{\sum_{i=1}^{k} d(p_i,q_i) \colon p_1 = x,\ q_k = y,\ k \in \mathbb{N}\},$$

where the infimum is taken over all choices of $\{p_i\}$ and $\{q_i\}$ such that q_i is R-equivalent to p_{i+1} for all $i = 1, \dots, k-1$. As usual, we associate with a semi-metric space (X, d_R) a metric space $(X/d_R, d_R)$ by identifying points with zero d_R-distances (and keep the same notation d_R for this metric; see Section 1.1). This space is called the *quotient metric space*. One also says that it results from *gluing* the space (X,d) along the relation R.

Exercise 3.1.13. Prove that d_R is indeed a semi-metric, i.e., it is nonnegative, symmetric and satisfies the triangle inequality.

We will consider quotients of length spaces only. It is easy to see that gluing a length space yields a length space. First of all, observe that $d_R \le d$, i.e., the gluing procedure never increases distances between (equivalence

classes of) points. Hence every d-continuous curve is d_R-continuous, and the d_R-length is not greater than the d-length. Now if $\{p_i\}_{i=1}^k$ and $\{q_i\}_{i=1}^k$ are as in the definition of $d_R(x,y)$, one can construct a curve connecting x and y in $(X/d_R, d_R)$ whose length is almost equal to $\sum d(p_i,q_i)$. To do this, simply concatenate almost shortest paths (of the original metric) connecting p_i to q_i, $i = 1,\dots,k$. Since q_i and p_{i+1} are identified in X/d_R, this curve is continuous in $(X/d_R, d_R)$. Its length can be made arbitrarily close to $\sum d(p_i,q_i)$, and hence to $d_R(x,y)$. Therefore $(X/d_r, d_R)$ is a length space.

Let us make a few immediate observations. First of all, if two points are R-equivalent, then they obviously get identified when passing to X/d_R. On the other hand, it is possible that the relation $d_R = 0$ defining the quotient X/d_R is actually stronger than R. This means that more points are identified in X/d_R than in X/R. In particular, $(X/d_R, d_R)$ is *not* necessarily homeomorphic to the topological quotient X/R. To see how this can happen, glue together all rational points in $[0,1]$. The topological quotient is a very wild (even nonmetrizable) space, while the metric quotient is just a point. The same effect takes place in Example 3.1.10 above.

Even if no additional identifications occur by metric reasons (i.e., X/d_R and X/R coincide as sets), still the topology of the metric quotient may differ from the topological quotient (see Example 3.1.17 below).

Exercise 3.1.14. Suppose that X/d_R and X/R coincide as sets, i.e., the relation $d_R = 0$ coincide with R. Prove that

1. The topology of the metric quotient X/d_R is weaker than that of the topological quotient X/R.

2. If X is compact, then the two topologies coincide.

Hint: Every continuous bijection from a compact topological space to a Hausdorff space is a homeomorphism. Observe that X/R is compact, X/d_R is Hausdorff, and the identity map from X/R to X/d_R is continuous.

The procedure of passing to a metric quotient can be used to glue several metric spaces together. First we need a definition:

Definition 3.1.15. Let (X_α, d_α) be a collection of length spaces. Consider the disjoint union $\cup_\alpha X_\alpha$. We introduce a length metric d on this disjoint union defining $d(x,y)$ by the following rule:

If $x,y \in X_\alpha$ for some α, then $d(x,y) = d_\alpha(x,y)$;

Otherwise, $d(x,y) = \infty$. The metric d will be referred to as the *length metric of disjoint union.*

Now let us begin with the simplest example of two length spaces (X, d_X) and (Y, d_Y) and a bijection $I : X' \to Y'$ between subsets $X' \subset X$ and

$Y' \subset Y$. Consider the disjoint union $Z = X \sqcup Y$ of X and Y together with the length metric of disjoint union. To glue X and Y along I, we first introduce the equivalence relation R on Z generated by the relations $x \sim y$ if $I(x) = y$. Now the result of gluing X and Y along I by definition is the quotient space $(Z/d_R, d_R)$.

More generally, if we want to glue a collection of length spaces along an equivalence relation defined on their disjoint union, we first endow this disjoint union with the length metric of disjoint union and then pass to the metric quotient.

Numerous examples of gluing will appear in this chapter and throughout the book. Here we give a few first examples:

Example 3.1.16. Consider the segment $I = [0,1]$ with its canonical metric d. Introduce the equivalence relation R generated by $0 \sim 1$. It is easy to see that $(I/d_R, d_R)$ is isometric to a circle of length 1.

Example 3.1.17. This example may seem as elementary as the first one; however most of the students who see it for the first time make various mistakes when analyzing its topology. Consider a countable collection of disjoint intervals I_i and identify together all their left ends. The space resulting from this *metric* gluing looks like a point with countably many needles sticking out.

Question. Is this a compact space?

Answer. The topology of this space depends on the lengths of intervals I_i!

In particular, if all lengths of I_i are equal to 1, this space is homeomorphic to the topological quotient: this is the bouquet of countably many segments. In particular, this is a noncompact space: the right ends of intervals form a countable collection of points with all pairwise distances equal to 2. On the other hand, if the length of I_i is $1/i$, the "metric bouquet" is compact (prove this!). Since compactness is a topological property, we see that this gluing leads to different topologies resulting from the same topological spaces with the same equivalence relation factored out.

Exercise 3.1.18. Can you find all topological spaces that may result from this example for different lengths of I_i?

Example 3.1.19. Begin with the square $Q = [0,1] \times [0,1]$ with its Euclidean metric. Introduce the equivalence R generated by the relations of the form $(0,x) \sim (1,x)$ and $(x,0) \sim (x,1)$, $x \in [0,1]$. The quotient space is a torus. It may seem that, while nothing happens to the interior points of Q, the boundary points and especially the vertices (all the four glued together) are somewhat distinguished and the quotient metric may look different near to these points. This is not the case:

Exercise 3.1.20. Prove that the (equivalence class of) a vertex of Q in the quotient metric has a neighborhood *isometric* to a Euclidean region.

This exercise explains why this quotient space is called a flat torus.

Example 3.1.21. Let G be a group of isometries (rigid motions) of $\mathbb{R}^2$. Introduce an equivalence relation R defined as follows: xRy iff there exists a $g \in G$ such that $x = gy$ (i.e., x and y belong to the same orbit).

Exercise 3.1.22. Prove that the quotient metric between (the equivalence classes of) two points x and y is equal to $\inf_{g\in G} |x - g(y)|$, where $|\cdot|$ stands for Euclidean distance.

Different choices of G lead to various examples. For instance, the group of translations by vectors with integer coordinates leads to the same torus as in the previous example. For G being the whole group of rotations, the quotient space is isometric to the ray $[0, \infty)$. The reader may wish to play with the examples of G generated by

(a) all translations along the x-axis;

(b) the transformation $(x, y) \to (x + 1, -y)$;

(c) the translation by vector $(0, 1)$ and the symmetry in the y-axis;

(d) the transformations $(x, y) \to (x + 1, -y)$ and $(x, y) \to (-x, y + 1)$.

3.1.3. Maximal metric. This section introduces a very useful construction that generalizes (and in a sense axiomatizes) the procedure of gluing. To understand the punch line of the construction, let us informally formulate the properties that a gluing procedure must definitely possess; we will see that it is uniquely determined by these properties.

Informal axiomatizing of gluing. Instead of saying that gluing makes one point out of two (or more) points, we could say that we glue two points "metrically" by setting zero distance between them. This is certainly impossible since our definition of distance says that the distance between different points is positive. This is, however, a minor difficulty: we relaxed this assumption and defined *semi-metrics*, which will perfectly suit our purpose. There is a more serious problem: just assigning some distances to be zero and keeping other distances unchanged may violate the triangle inequality, and this is what we cannot afford to sacrifice. This forces us to modify other distances at the same time. On the other hand, it would be very unnatural if gluing some points together could *increase* any distance. We also want to decrease distances only to the extent dictated by gluing. Summarizing this, we informally *define the metric resulting from gluing as the largest metric not exceeding the original one and such that the distances between identified points are zero.*

Maximal metric. Now we can proceed with formal definitions. To formalize the words "the largest metric..." in the informal definition of gluing, we need the following lemma.

Lemma 3.1.23. *Let $b : X \times X \to \mathbb{R}_+ \cup \{+\infty\}$ be an arbitrary function. Consider the class D of all semi-metrics d on X such that $d \le b$, i.e., $d(x, y) \le b(x, y)$ for all $x, y \in X$. Then D contains a unique maximal semi-metric d_m such that $d_m \ge d$ for all $d \in D$.*

Proof. For every two points $x, y \in X$ define

$$d_m(x, y) = \sup\{d(x, y) : d \in D\}.$$

The function d_m is nonnegative, symmetric and satisfies $d_m \le b$. So all we have to prove is the triangle inequality for d_m. For all $x, y, z \in X$ one has

$$\begin{aligned} d_m(x, y) = \sup_{d\in D} d(x, y) &\le \sup_{d\in D}(d(x, z) + d(z, y)) \\ &\le \sup_{d\in D} d(x, z) + \sup_{d\in D} d(z, y) = d_m(x, z) + d_m(z, y), \end{aligned}$$

and the lemma follows. □

Corollary 3.1.24. *Let a set X be covered by a collection of subsets $\{X_\alpha\}$ each carrying a semi-metric d_α. Consider the class D of all semi-metrics d (possibly taking infinite values) such that $d(x, y) \le d_\alpha(x, y)$ whenever $x, y \in X_\alpha$. Then D contains a unique maximal semi-metric d_m such that $d_m(x, y) \ge d(x, y)$ for all $d \in D$ and $x, y \in X$. If all d_α are intrinsic (semi-) metrics, then so is d_m.*

Proof. One may assume that the metrics d_α are defined on the whole space X by setting $d_\alpha(x, y) = \infty$ if $x \notin X_\alpha$ or $y \notin X_\alpha$. To prove the existence of d_m, apply Lemma 3.1.23 to the function $b(x, y) = \inf_\alpha d_\alpha(x, y)$. To prove that d_m is intrinsic if d_α are, consider the intrinsic metric $\widehat{d}_m$ induced by d_m. Since all d_α are intrinsic and $d_m \le d_\alpha$, it follows that $\widehat{d}_m \le d_\alpha$, so $\widehat{d}_m$ belongs to D. Then $\widehat{d}_m = d_m$ because d_m is maximal. □

An important case is a topological space X covered by a collection of metric spaces (X_α, d_α) whose metrics agree on the intersections of their domains and are consistent with the topology (compare with Lemma 3.1.1). If X is connected, the sets X_α are open and the semi-metrics d_α are finite, then the maximal semi-metric is finite.

Exercise 3.1.25. Prove this statement.

Exercise 3.1.26. Let (X, d) be a metric space and set d_ε to be the maximum metric w.r.t. the covering of X by all balls of radius ε equipped with the restrictions of d. Let $d_0(x, y) = \sup_{\varepsilon>0} d_\varepsilon(x, y)$. Prove that if d is complete, then d_0 is the intrinsic metric induced by d.

Now we are ready to give a definition of gluing in terms of maximal metrics.

Theorem 3.1.27. *Let (X, d) be a metric space and R an equivalence relation on X. Consider the function b_R on $X \times X$ defined by*

$$b_R(x,y) = \begin{cases} 0 & \text{if } x \text{ is } R\text{-equivalent to } y, \\ d(x,y) & \text{otherwise.} \end{cases}$$

Then the maximal semi-metric among those not exceeding b_R coincides with the semi-metric d_R obtained by gluing (X, d) along R as in Definition 3.1.12.

Proof. Let D denote the class of semi-metrics not exceeding b_R. Obviously $d_R \in D$, so it suffices to prove that $d_R \geq d'$ for any semi-metric $d' \in D$. Let $x, y \in X$ and $\{p_i\}_{i=1}^k$, $\{q_i\}_{i=1}^k$ be as in Definition 3.1.12. Then by the triangle inequality

$$d'(x,y) \leq \sum_{i=1}^{k} d'(p_i, q_i) + \sum_{i=1}^{k-1} d'(q_i, p_{i+1}) \leq \sum_{i=1}^{k} d(p_i, q_i)$$

since $b_R(p_i, q_i) \leq d(p_i, q_i)$ and $b_R(q_i, p_{i+1}) = 0$. Hence $d' \leq d_R$. □

Note that the function b_R in the theorem is the minimum of two length semi-metrics: the original metric d and the one which equals 0 for R-equivalent points and ∞ otherwise. Thus we are in conditions of Corollary 3.1.24, which gives us another proof for the fact that d_R is intrinsic.

3.2. Polyhedral Spaces

3.2.1. Polyhedral metrics.

Two-dimensional finite polyhedral spaces. Roughly speaking, a polyhedral space is a length space that can be obtained by using convex polyhedra as building blocks and a collection of isometries between their faces indicating how these blocks should be attached to each other.

Two-dimensional polyhedral spaces are glued from points, segments and polygons. Since every polygon can be cut into several triangles by its diagonals, we can do with triangles instead of polygons.

Let us define 0-, 1- and 2-dimensional finite polyhedral spaces.

A 0-dimensional finite polyhedral space P_0 is just a space consisting of finitely many points, with all pairwise distances being infinite. These points are called *vertices*.

To construct a 1-dimensional finite polyhedral space, we begin with a 0-dimensional finite polyhedral space P_0 and a finite collection of segments $E = \{E_i\}$ (each viewed as a length space isometric to a segment). For each

E_i, fix an injective map e_i sending the endpoints of the segment E_i to P_0. Gluing E to P_0 along $\{e_i\}$ produces a 1-dimensional polyhedral space P_1. Note that different points from the same segment from E never glue together and hence we can regard segments from E lying in P_1. These segments are called *edges*. 1-dimensional polyhedral spaces are called *graphs* (see the next section).

To construct a 2-dimensional polyhedral space, begin with a 1-dimensional polyhedral space P_1 and take a finite collection of polygons $F = \{F_i\}$. For each polygon F_i fix an injective map f_i from the boundary of F_i to P_1 such that f_i bijectively maps each side of F_i onto an edge from E of the same length. Moreover f_i must map sides of the polygon to the respective edges *isometrically* in the sense that lengths of subintervals are preserved. Gluing F to P_1 along $\{f_i\}$ produces a 2-dimensional polyhedral space P_2. Copies of F_i lying in P_2 are called *faces*.

Several examples are given as the following exercises.

Exercise 3.2.1. Prove that the surface of a convex polyhedron in $\mathbb{R}^3$ is a 2-dimensional polyhedral space.

Exercise 3.2.2. Prove that the flat torus from Example 3.1.19 is a 2-dimensional polyhedral space.

Exercise 3.2.3. Prove that each two-dimensional finite polyhedral space can be glued in such a way that the distance between two points in the same face can be measured by a segment in this face. In other words, there is a shortest path that does not traverse other faces.

General polyhedral spaces. Here we introduce general polyhedral spaces. We do not impose such restrictions as local finiteness, finite dimension and even local compactness. Such spaces can be quite useful, and we will give a few examples of polyhedral monsters. Most of other examples of polyhedral spaces that arise further in this book are two-dimensional locally-finite (or, better, finite) polyhedral spaces. Hence, this section can be omitted for the first reading.

We will not repeat standard definitions from the theory of convex polyhedra in Euclidean space. The reader who is not comfortable with higher-dimensional convex polyhedra can think of two-dimensional polygons and perhaps three-dimensional polyhedra.

Since every polyhedron can be triangulated, it is actually enough to work with simplexes. One-dimensional simplexes are segments, two-dimensional simplexes are triangles and three-dimensional simplexes are tetrahedra. Each tetrahedron has four triangular faces of dimension 2 and six one-dimensional faces (edges). Vertices may be regarded as 0-dimensional faces. The entire tetrahedron can be viewed as its three-dimensional face.

Let us formally define a polyhedral space.

Definition 3.2.4. Let a topological space (P, d) be covered by a collection of length spaces (P_α, d_α), each isometric to a simplex (or a convex polyhedron). Assume that for any two simplexes P_α and P_β their intersection $P_\alpha \cap P_\beta$ is a face in both of them, and that the restrictions of the metrics of P_α and P_β to $P_\alpha \cap P_\beta$ coincide. Consider the maximal metric d majorized by all metrics d_α. The length space (P, d) is called a *polyhedral space*, or, more precisely, a *Euclidean polyhedral space*. (The latter term stresses the fact that P is built from Euclidean polyhedra. One may as well consider spaces glued from simplexes equipped with some non-Euclidean length metrics.)

The polyhedra P_α are called *faces* of P. 0-dimensional and 1-dimensional faces are called *vertices* and *edges*, respectively. A particular representation of a polyhedral space as a union of faces P_α is referred to as a *triangulation*. We will usually assume that every polyhedral space comes with some fixed triangulation.

There is a more constructive reformulation of this definition. Namely, we can begin with a collection of simplexes (convex polyhedra) $\{P_\alpha\}$ and a collection of isometries $\{I_\delta\}$, such that each isometry $I_\delta : P_\alpha^f \to P_\beta^f$ maps a face $P_\alpha^f \subset P_\alpha$ onto a face $P_\beta^f \subset P_\beta$ for some P_α and P_β. Form the disjoint union $Z = \bigcup_\alpha P_\alpha$ and introduce an equivalence relation R on Z generated by the relations $x \sim I_\delta(x)$.

The reader should be already suspecting that what we want to do next is to glue Z (endowed with the length metric of disjoint union) along the relation R. There is, however, a little formal problem here. It may happen that the relation R identifies two different points of the same simplex P_α. If we want this construction to be coherent with the previous definition, this should be prohibited. Thus, the quotient space (P, d) resulting from gluing Z along the relation R is a polyhedral space if each P_α projects to P injectively. Fortunately this requirement does not restrict the class of spaces that can be produced this way:

Exercise 3.2.5. Prove that every space P glued from convex polyhedra by means of isometries of their faces is a polyhedral space.

Hint: Subdivide polyhedra into smaller ones.

Notice that an isometry between two simplexes (as well as convex polyhedra) is uniquely determined by the images of the vertices. Thus the data needed to describe a polyhedral space are more or less combinatorial. More precisely, to define a polyhedral space it suffices to fix the lengths of edges of simplexes P_i (to determine their geometry) and give a combinatorial scheme of gluing.

By the *dimension* of a polyhedral space we mean the maximal dimension of polyhedra used to glue it. A polyhedral space is said to be *dimensionally homogeneous* is it can be glued from polyhedra of the same dimension. Equivalently, a polyhedral space P is dimensionally homogeneous if every point of P belongs to an n-dimensional face where n is the dimension of P.

A polyhedral space is said to be *finite* if is has finitely many faces, and *locally finite* if every point possesses a neighborhood which intersects only finitely many faces.

Exercise 3.2.6. Prove that a polyhedral space P is locally finite if and only if every point of P belongs to finitely many faces.

For nice examples, we refer to the previous subsection. Here we present a couple of a bit terrifying examples.

Example 3.2.7. Consider an infinite sequence of cubes $P_i = [0,1]^i$. There are the isometries $I_i : P_{i-1} \to P_i$ acting by adding 0 for the last coordinate:

$$I_i(x_1, x_2, \dots x_{i-1}) = (x_1, x_2, \dots x_{i-1}, 0).$$

Thus we can form a polyhedral space. This is an infinite-dimensional cube of all finite sequences of reals between 0 and 1.

Exercise 3.2.8. Can you define an infinite-dimensional simplex?

Example 3.2.9. Let (X, d) be a length space. Choose a positive ε. For each finite collection of points $A = \{x_1, \dots, x_k \in X\}$ such that all pairwise distances $d(x_i, x_j)$ are bounded by ε ($d(x_i, x_j) \le \varepsilon$), form a simplex σ_A with vertices $v_{x_1}, v_{x_2}, \dots, v_{x_k}$ indexed by the points in A and such that the lengths of edges $|v_{x_i} v_{x_j}|$ equal the corresponding distances $d(x_i, x_j)$. (If these length do not determine a valid simplex, just ignore this collection of points.) For two simplexes σ_A and σ_B, with $A = \{x_1, \dots, x_k \in X\}$ and $B = \{y_1, \dots, y_l \in X\}$, we consider the intersection $A \cap B$. If this intersection is empty, there is no gluing rule for these simplexes. Otherwise, the vertices indexed by points from this intersection determine a face of each simplex, and we glue the simplexes by the isometry that maps each vertex of this face to the vertex with the same index. (Recall that every collection of vertices determines a face of a simplex, and an isometry between simplexes is determined by images of the vertices.) Denote the resulting polyhedral space by $P_\varepsilon(X)$.

This polyhedral space naturally contains an embedded copy of X. Indeed, all the vertices indexed by the same point $x \in X$ get identified according to our gluing rule, and this vertex v_x represents x in $P_\varepsilon(X)$.

Exercise 3.2.10. Show that the embedding $x \to v_x$ is an isometry onto its image; in other words, $P_\varepsilon(X)$ contains an isometric copy of X.

When ε approaches zero, polyhedral spaces $P_\varepsilon(X)$ in a certain sense get closer and closer to X (formally, $P_\varepsilon(X)$ converges to X in Gromov–Hausdorff topology; see Chapter 7). The spaces $P_\varepsilon(X)$ are often called *polyhedral regularizations* of X.

3.2.2. Metric graphs. We presume that the reader is familiar with topological graphs at least as a nice way to draw a few cities connected by several roads. Metric graphs are just one-dimensional polyhedral spaces. Regardless of the fact that the definition of a polyhedral space is given in the previous section, we repeat it here to better visualize its meaning in this simplest case. This definition generalizes the example of a "cobweb" from the first chapter.

By a metric segment (of length a) we mean a metric space isometric to a segment $[0, a]$.

Definition 3.2.11. A (metrized) graph is the result of gluing of a disjoint collection of metric segments $\{E_i\}$ and points $\{v_j\}$ (regarded with the length metric of disjoint union) along an equivalence relation R defined on the union of the set $\{v_j\}$ and the set of the endpoints of the segments.

In other words, we consider the maximal semi-metric that is bounded by all metrics of segments and an additional semi-metric $d_R(x, y)$ that is zero if xRy and infinite otherwise, and then identify points that are within zero distance from each other w.r.t. the maximal semi-metric.

The segments $\{E_i\}$ are called edges, and the equivalence classes of the endpoints are called vertices of the graph. The length of an edge is just the length of the segment (it may be different from the distance between its endpoints in the graph). The cardinality of an equivalence class representing a vertex is called its *degree*. Most of the graphs considered in this book will be locally finite and therefore with finite degree of every vertex (one exception is a monster-graph in an example in this section).

The most natural way to define a graph is to take a set V of vertices, indicate which pairs of vertices are connected by edges, and specify lengths of these edges. To recover the original definition from these data, take a collection of segments $\{E_i\}$ corresponding to the desired edges and having the given lengths. Then, if the edge corresponding to E_i should connect vertices a and b, let one endpoint of E_i be equivalent to a and another be equivalent to b. This generates an equivalence relation R used to glue the graph from V and $\{E_i\}$.

In particular, every topological graph can be turned to a metric graph by assigning lengths to its edges. However, one should remember that (for infinite graphs) this may change the topology of the space. See also Exercise 3.2.14 below.

Note that a pair of vertices may be connected by two or more different edges. Definition 3.2.11 even allows a graph to have loops, that is, edges connecting a point to itself. One can always obtain a graph without loops and multiple edges by dividing edges into smaller ones (compare with Exercise 3.2.5).

To get better understanding of the definitions, we suggest the following

Exercise 3.2.12. Unwind the definition of gluing and verify that for finite graphs our definition agrees with the usual one: the distance from x to y is the shortest way of reaching y from x by means of traversing finitely many segments such that the terminating point of each segment is identified with the beginning of the next one. Is this true for infinite graphs? What if we replace "the shortest" by "the infimum"?

Caution: As well as in other cases of gluing, points that are nonequivalent w.r.t. R can still get identified.

Exercise 3.2.13. Give such an example.

Exercise 3.2.14. Prove that additional identifications cannot occur if the graph is locally finite, or if the lengths of all edges are bounded from below by a positive constant. Moreover in these cases the topology of the metric graph coincides with that of the topological graph (i.e., of the topological quotient of the disjoint union by the same equivalence relation).

On the other hand, the identification $X \to X/d_R$ can never affect the interiors of segments:

Exercise 3.2.15. Prove that points in the interiors of edges do not get identified and thus the disjoint union of these interiors is embedded into the graph.

In other words, every edge is a simple curve connecting two vertices and not passing through other vertices, and the interiors of different edges have no common points.

Exercise 3.2.16. Prove that the length of every edge as a curve in the graph equals the length of the segment from which the edge was obtained.

Exercise 3.2.17. Let X be a length space and $V \subset X$ a finite set. Suppose that $X \setminus V$ is covered by a finite collection of rectifiable curves, each connecting two points of V, and these curves have no self-intersections and do not intersect one another (except at endpoints). Prove that X is a metric graph.

Exercise 3.2.18. Show that the statement of the previous exercise fails without the assumption that the collection of curves is finite.

Hint: Every length space is a union of points, but not all length spaces are 0-dimensional polyhedra.

Although our definitions sound quite scientific, the object is very familiar and our everyday intuition is a good guide through simple cases of graphs. For instance, to find intrinsic distances and shortest curves in the frame of a cube is a good exercise for a 10-year old kid. Graphs may even seem dull as being too discrete and combinatorial for a geometrically-thinking reader. To relieve this feeling we suggest the following approach: one can try to approximate an arbitrary length space by graphs whose vertices are dense enough in the space; see Chapter 7 for more details. For the most skeptical reader, the following example shows that every metric space can be realized as the set of vertices of a graph, with the distance between points being equal to that in the graph. Alas, this graph is apparently a hardly useful monster.

Example 3.2.19. Begin with a general metric space (X, d); consider a disjoint union of segments $I_{x,y}$ parameterized by pairs $x, y \in X$. Equip each segment $I_{x,y}$ with the metric of the segment $[0, d(x,y)]$. These are edges. The identifications are most natural: one identifies the right ends of two segments $I_{x,y}$ and $I_{x',y'}$ if $y = y$ and the left ends if $x = x'$. We leave as an exercise to verify that the set X can be canonically identified with the set of vertices of the graph and that the intrinsic metric of this graph restricted to the set of its vertices is the original metric d in X. (To get better intuition behind this construction, one can think of it this way: for each pair of points $x, y \in X$, one takes a segment of length $d(x, y)$ and attaches one of its ends to x and the other one to y.)

3.2.3. Word metrics. There is an important class of graphs that arise from groups. Let G be a finitely generated group; that is, it contains a finite generating set: a collection of elements $S = \{g_1, \dots, g_k\} \subset G$ such that every element of G can be represented as a finite product of elements of S. We also require S to be symmetric: for every $s \in S$, s^{-1} is also in S.

Remark 3.2.20. If S is not symmetric, it can be symmetrized by including all inverses of its elements. Thus we will not list all inverses of elements in the generating sets, following a commonly accepted convention that all inverses are also included there by default.

Given a group G and a generating set $S = \{g_1, \dots, g_k\} \subset G$, one constructs the Cayley graph of (G, S). Its vertices are just elements of G, and two elements $g, h \in G$ are connected by an edge if and only if $gh^{-1} \in S$. In other words, every vertex $g \in G$ is connected by edges to the k vertices $g_1g, g_2g, \dots, g_kg$. The lengths of all edges are equal to 1.

In terms of Definition 3.2.11, the Cayley graph can be described as follows. Take a collection of unit intervals I_g^s labeled by the pairs ($g \in G, s \in S$), and glue them along the equivalence relation generated by the following identifications: if $h = sg$, identify the right end of the segment I_g^s with the left ends of segments I_h^t for all $t \in S$.

Exercise 3.2.21. What are the Cayley graphs for the following groups and generating sets:

1. G is $\mathbb{Z}$ generated by $\{1\}$;
2. G is $\mathbb{Z}/m\mathbb{Z}$ generated by one generator;
3. G is $\mathbb{Z}^2$ generated by $\{(0,1),(1,0)\}$;
4. G is $\mathbb{Z}^2$ generated by $\{(0,1),(1,0),(1,1)\}$;
5. G is a free group generated by two generators a, b.

Answers. 1. Real line made of unit segments.

2. A regular m-gone with side 1.

3. The grid on the plane formed by lines $x = n$ and $y = n$ for all integers n.

4. The grid on the plane formed by the lines $x = n$, $y = n$ and $x - y = n$ for all integers n. To make the picture more consistent with the metric (where all edges have unit length), draw a grid dividing the plane into regular triangles with unit sides.

5. An infinite tree with all nodes of degree 4. (A *tree* is a graph which contains no subsets homeomorphic to the circle.)

Definition 3.2.22. The restriction to G of the intrinsic distance associated with the length structure of a Cayley graph is called the *word metric* on G (with respect to a given set of generators).

Exercise 3.2.23. Check that this is a finite metric.

Exercise 3.2.24. Write down the word metrics and draw word metric's balls of radius 5 for the first four examples in Exercise 3.2.21.

Exercise 3.2.25. Show that the word distance between g and h is equal to the smallest number n such that g can be represented as $g = h_1 h_2 \dots h_n h$, with all $h_i \in S$ (recall that we assumed that S is symmetrical). That is why this distance is called word metric: the distance between two elements g and h is the length of the shortest word in generators that is equal to $g^{-1}h$.

The role of this construction in both combinatorial group theory and modern geometry is difficult to over-estimate; we will dwell on it a little further in Chapter 8.

3.3. Isometries and Quotients

Let us recall the notion of isometry:

Definition 3.3.1. Let X and Y be two metric spaces. A map $f : X \to Y$ is called an *isometry onto its image* if it preserves distance. The latter means that $|f(x)f(y)| = |xy|$ for any two points $x, y \in X$ (and thus it is automatically injective).

A map that is surjective and preserves distance is called an *isometry*. Two spaces are said to be *isometric* if there exists an isometry from one to the other.

It is clear that an isometry is automatically a homeomorphism and being isometric is an equivalence relation. It is clear that an isometry onto its image maps a curve to a curve of the same length and therefore it is an isomorphism of length structures. For an isometry, the image of a shortest path (geodesic, sphere, ball, etc.) is again a shortest path (geodesic, sphere, ball, etc.). We can say that metric geometry studies classes of isometric spaces.

Isometries of a space to itself obviously form a group (check this!) naturally called the *isometry group*. We will denote it by $Iso(X)$. For a "generic" space this group is trivial. (Can you give an example of a length space with trivial isometry group?) If the isometry group of a length space is nontrivial, this is often expressed by saying that the space possesses symmetries. There is a remarkable class of very symmetrical spaces, namely *homogeneous* spaces.

Definition 3.3.2. A length space X is said to be *homogeneous* if for every $x, y \in X$ there exists an isometry $I : X \to X$ such that $I(x) = y$.

Example 3.3.3. Euclidean spaces and (round) spheres are homogeneous spaces. The cylinder $x^2 + y^2 = 1$ in $\mathbb{R}^3$ and the torus from Example 3.1.19 are also homogeneous. On the other hand, the cone $x^2 + y^2 = z^2$ and the paraboloid $x^2 + y^2 = z$ are not homogeneous spaces.

Can you find the isometry groups for these spaces?

There is an important notion, which is relevant to introduce here.

Definition 3.3.4. A metric space X is said to be *locally isometric* to a homogeneous space Y if each point in X possesses a neighborhood isometric to an open set in Y. Spaces locally isometric to a Euclidean space are said to be *flat*.

Flat spaces cannot be distinguished from Euclidean space by local measurements: they look the same. For instance, the cylinder $x^2 + y^2 = 1$ in $\mathbb{R}^3$

is flat. To realize that this is a cylinder, a two-dimensional creature would have to undertake a long "Magellan" trip around the cylinder. The cone $x^2 + y^2 = z^2, z \geq 0$, is not flat, as it looks very different at its apex than the Euclidean plane.

Question: What would you suggest for a two-dimensional creature living in the cone to tell that this is not a two-plane? What about a two-dimensional creature living in the sphere? Did we have to travel around the Earth to figure out that it is not planar? Notice that the experiments with a horizon are illegal in this set-up: they use measurements in the ambient space, and we are allowed to use intrinsic measurements only. The answer to these questions is not that easy, but the reader should be able to give them after reading this textbook.

Quotient spaces. In this section we are going to deal with a subgroup of an isometry group. The reader who is familiar with group actions should rather think of a group G acting on X by isometries. We recall the definition of group action here:

Definition 3.3.5. One says that a group G acts on a set X if there is a map $\varphi : G \times X \to X$, which we abbreviate as $\varphi(g, x) = g(x)$, such that

(i) $gh(x) = g(h(x))$ and

(ii) $e(x) = x$

for every $g, h \in G$, $x \in X$. Here e is the unit of G.

Consider a subgroup $G \subset Iso(X)$ of the isometry group of a length space (X, d). Introduce an equivalence relation R_G: points x and y in X are equivalent iff $x = g(y)$ for some $g \in G$ (check that this is an equivalence relation!)

Since we have an equivalence relation on a length space, we can glue together equivalent points and consider the quotient metric on the space resulting from this identification as in Definition 3.1.12 of gluing.

It turns out that, for an equivalence relation arising from an isometry subgroup, the space resulting from the gluing construction is just X/R_G (usually denoted by just X/G) and the distance between two points in X/G is the infimum of distances between their representatives in X.

More formally, for $\overline{x}, \overline{y} \in X/G$, set

$$\overline{d}(\overline{x}, \overline{y}) = \inf\{d(x, y) : \quad x \in \overline{x}, y \in \overline{y}\}$$

for every $\overline{x}, \overline{y} \in X/G$.

Recall that the equivalence class of a point x is called the orbit $O(x) = \{g(x) : g \in G\}$. Then the definition of $\bar{d}$ can be re-formulated as follows:

$$\bar{d}(O(x), O(y)) = \inf_{g \in G} d(x, g(y)).$$

Lemma 3.3.6. *$\bar{d}$ coincides with the quotient metric d_{R_G}.*

Proof. By the definition of gluing, the distance between the classes of $p, q \in X$ in the quotient metric is the infimum of sums

$$\sum_{i=0}^{k} d(p_i, q_i)$$

for all finite collections of $\{p_i, q_i, i = 0, 1, \dots, k, k \in \mathbb{N}\}$ such that $p_0 = p$, $q_k = q$ and with q_i being equivalent to p_{i+1} for all i. The latter means that there are isometries $g_i \in G$ such that $g_i(q_i) = p_{i+1}$. We can use these isometries to assemble all segments p_i, q_i together. Namely, consider a new sequence of points

$$\begin{aligned} \tilde{p}_0 = p_0 = p, &\quad \tilde{q}_0 = q_0, \\ \tilde{p}_1 = g_1(p_1), &\quad \tilde{q}_1 = g_1(q_1), \\ \tilde{p}_2 = g_1(g_2(p_2)), &\quad \tilde{q}_2 = g_1(g_2(q_2)), \\ &\dots \\ \tilde{p}_k = g_1 \circ g_2 \circ \dots \circ g_k(p_k), &\quad \tilde{q}_k = g_1 \circ g_2 \circ \dots \circ g_k(q_k). \end{aligned}$$

Since all g_i and thus their compositions $g_1 \circ g_2 \circ \dots \circ g_i(q_i)$ are isometries, $d(p_i, q_i) = d(\tilde{p}_i, \tilde{q}_i)$ and thus

$$\sum_{i=0}^{k} d(p_i, q_i) = \sum_{i=0}^{k} d(\tilde{p}_i, \tilde{q}_i).$$

On the other hand, by the choice of isometries g_i, $\tilde{q}_i = \tilde{p}_{i+1}$. Thus

$$d(x, \tilde{q}_k) \le \sum_{i=0}^{k} (d(\tilde{p}_i, \tilde{q}_i) + d(\tilde{q}_i, \tilde{p}_{i+1})) = \sum_{i=0}^{k} d(p_i, q_i).$$

Recalling that

$$\tilde{q}_k = g_1 \circ g_2 \circ \dots \circ g_k(q_k) = g_1 \circ g_2 \circ \dots \circ g_k(q)$$

and thus $\tilde{q}_k$ belongs to the orbit $O(q)$ of q, we conclude that the distance between the orbits of p and q is less than or equal to the distance between the equivalence classes of p and q in the quotient metric. The opposite inequality is obvious. Indeed, if $q' \in O(q)$ and $p' \in O(p)$, then p' and q' are equivalent to p and q respectively, and hence one chooses the path $p_0 = q_0 = p$, $p_1 = p'$, $q_1 = q'$ and $p_2 = q_2 = q$ with only one nonzero jump, and the length of this path is equal to $d(p', q')$. □

Exercise 3.3.7. Give a direct proof (that is, without referring to Definition 3.1.12 of gluing and the notion of maximal metric) that the metric $\bar{d}$ defined this way is an intrinsic metric.

Hint: Here is a proof for the case when X is locally compact and complete and all orbits are compact. To get a general proof, one has to use approximations.

Given $a, b \in X/G$, choose their representatives $x \in a, y \in b$ (as a and b are equivalence classes) such that $\bar{d}(a,b) = d(x,y)$. This is possible since we assumed that all orbits are compact. Let z be a mid-point for x, y, i.e. $d(x,z) = d(y,z) = \frac{1}{2}d(x,y)$. Obviously $d(a, O(z)) \le d(x,z) = \frac{1}{2}d(x,y) = \frac{1}{2}d(a,b)$. Similarly $d(b, O(z)) \le \frac{1}{2}d(a,b)$. The triangle inequality shows that both inequalities turn out to be equalities, so $O(z)$ is a mid-point for a, b. □

Example 3.3.8. For the subgroup G of isometries of $\mathbb{R}^2$ acting by integer translations ($g(x,y) = (x+k, y+l)$, k, l are integers), the quotient space $\mathbb{R}^2/G$ is a torus (isometric to the flat torus from Example 3.1.19).

Example 3.3.9. For the subgroup G of isometries of $\mathbb{R}^2$ consisting of two elements, the identity and the symmetry $-\mathrm{id}\colon (x,y) \to (-x,-y)$, the quotient space is isometric to a cone $x^2 + y^2 = cz^2$, $z \ge 0$. Can you find this coefficient c? What about G being a finite group of rotations?

Example 3.3.10. For the subgroup G of isometries of $\mathbb{R}^2$ consisting of two elements, the identity and the symmetry $(x,y) \to (x,-y)$, the quotient space is a half-plane.

The reader may notice a difference between these examples: while the torus looks "the same at every point", both cones and half-planes have distinguished "singular" points (the vertex of a cone and the edge of a half-plane). In the next subsection we will see that these singular points come from fixed points of elements from G.

3.4. Local Isometries and Coverings

3.4.1. Local isometries.

Definition 3.4.1. A map $f : X \to Y$ is called a *local isometry* at $x \in X$ if x has a neighborhood U_x such that (the restriction of) f maps U_x isometrically onto an open set U_y in Y. A map which is a local isometry at every point is called a *local isometry*.

Notice that in this definition we consider U_x and U_y with the restrictions of the metrics; we do not induce a new length structure in them. This remark is important for the spaces where points may have no "convex" neighborhoods.

A local isometry can change distances. For example, it can map two distinct points into one point. However, every local isometry of a length space is a *nonexpanding map*; i.e., it never increases distances between points.

Exercise 3.4.2. Prove this statement.

Example 3.4.3. The map $f : \mathbb{R} \to S$ given by $f(t) = (\sin(t), \cos(t))$ is a local isometry (w.r.t. the intrinsic (angular) metric on S).

Example 3.4.4. More generally, if $f : X \to Y$ is a local homeomorphism and Y is a length space, then X admits a unique structure of a length space such that f is a local isometry. This is the length structure induced by f (see section 2.2 for the definition).

Exercise 3.4.5. Prove the statement formulated in this example.

The following exercise explains why the torus $\mathbb{R}^2/\mathbb{Z}^2$ is flat and the cone $\mathbb{R}^2/\{id, -id\}$ is not.

Exercise 3.4.6. Let a group G act by isometries on a length space X. (The reader still may think of G as just a subgroup of $Iso(X)$.) Assume that all orbits are discrete. Prove that the projection map $p : X \to X/G$ is a local isometry at a point x if and only if x is not fixed by a nonidentity element of G.

3.4.2. Covering maps. Covering maps are an important class of local homeomorphisms. When length spaces are considered, covering maps provide an important class of local isometries. Here we recall basic properties of covering maps and consider relations between covering maps and length metrics. As for topological properties, we only recall the definitions and formulate the most important facts. The proofs can be found in many books (see [**Mas**]).

We want to stress additional properties arising when covering maps of length spaces are considered.

Coverings in general. Let X and Y be topological spaces and $f\colon X \to Y$ a continuous map. An open set $V \subset Y$ is said to be *evenly covered* if its inverse image $f^{-1}(V)$ is a disjoint union of sets $U_i \subset X$ such that the restriction of f onto each U_i is a homeomorphism from U_i to V. The map f is a *covering map*, or simply *covering*, if every point $y \in Y$ has an evenly covered neighborhood. The space Y is called the *base* of the covering and X the *covering space*. Classical examples of covering maps include: the map $f\colon \mathbb{R} \to S^1$ given by $f(x) = (\cos x, \sin x)$; the standard covering of the torus by the plane, i.e., the map $F\colon \mathbb{R}^2 \to T^2 = S^1 \times S^1$ given by $F(x,y) = (f(x), f(y))$ where $f(x) = (\cos x, \sin x) \in S^1$; the projection from

the sphere S^2 to the projective plane $\mathbb{RP}^2$ (recall that $\mathbb{RP}^2$ is the quotient of S^2 by the equivalence relation $x \sim -x$).

Every covering map is a local homeomorphism. The converse is not true; for example, consider the inclusion map from the interval $(0,1)$ to $\mathbb{R}$.

If X and Y are connected and $f\colon X \to Y$ is a covering map, then the cardinality of $f^{-1}(y)$ does not depend on $y \in Y$ and is called the *number of sheets* of f (this "number" may be infinity). We consider covering maps of arcwise connected spaces only.

Suppose that $f\colon X \to Y$ is a covering map and Y is a length space. Since f is a local homeomorphism, there is a unique length metric on X such that f is a local isometry. This metric on X is called the *lift* of the metric of X.

For a reader already familiar with smooth manifolds and Riemannian metrics, we note that if Y is a Riemannian manifold, then its differential structure and Riemannian metric can be similarly lifted to X.

Exercise 3.4.7. Let X and Y be length spaces, and let $f\colon X \to Y$ be a covering map and a local isometry. Let $y \in Y$ and a ball $B_r(y)$ be an evenly covered neighborhood. Prove that the distances between distinct points of $f^{-1}(y)$ are no less than $2r$.

Exercise 3.4.8. Let X and Y be length spaces, and let $f\colon X \to Y$ be a covering map and a local isometry, and suppose that Y is complete. Prove that X is complete too.

Hint: Since the map is nonexpanding, it maps a Cauchy sequence in X to a Cauchy sequence in Y. Consider the limit of the latter and an evenly covered neighborhood of it.

Exercise 3.4.9. Let X and Y be length spaces, and let $f\colon X \to Y$ be a covering map and a local isometry, $y_1, y_2 \in Y$, $\varepsilon > 0$. Prove that for every $x_1 \in f^{-1}(y_1)$ there exists an $x_2 \in f^{-1}(y_2)$ such that $|x_1x_2| \le |y_1y_2| + \varepsilon$. If the metric of Y is strictly intrinsic, then the same is true with $\varepsilon = 0$.

Hint: Utilize the standard topological lemma about lifting curves: for every curve $\gamma\colon [a,b] \to Y$ and every $x \in f^{-1}(\gamma(a))$ there exists a (unique) curve $\tilde{\gamma}\colon [a,b] \to X$ such that $f \circ \tilde{\gamma} = \gamma$ and $\tilde{\gamma}(a) = x$.

Exercise 3.4.10. Prove the converse to Exercise 3.4.8. Namely if X and Y are length spaces, $f\colon X \to Y$ is a covering map and a local isometry, and X is complete, then Y is complete too.

Hint: For a given Cauchy sequence $\{y_n\}$ in Y construct a Cauchy sequence $\{x_n\}$ in X such that $f(x_n) = y_n$.

Universal covering. A covering map $f\colon X \to Y$ is called a *universal covering* if X is simply connected. In this case, X is called a *universal covering space* for Y.

The next theorem tells that every "not too weird" topological space possesses a unique universal covering. To make this more precise, we give some definitions first. Two covering maps $f_1\colon X_1 \to Y$ and $f_2\colon X_2 \to Y$ are said to be *equivalent* if there is a homeomorphism $h\colon X_1 \to X_2$ such that $f_1 = f_2 \circ h$. It is natural not to distinguish equivalent coverings.

A topological space Y is *locally arcwise connected* if for every point $y \in Y$ and every neighborhood U of y there is a smaller neighborhood U', $y \in U' \subset U$, such that every two points $a, b \in U'$ can be connected by a path in U. A space Y is *locally simply connected in the large* or semi-locally simply connected if every point $y \in Y$ has a neighborhood U such that every loop contained in U is contractible in Y. (In other words, the image of the fundamental group $\pi_1(U, y)$ under the homomorphism $\pi_1(U, y) \to \pi_1(Y, y)$ induced by the inclusion of U to Y is trivial.)

Theorem 3.4.11. *If a topological space Y is a connected, locally arcwise connected and locally simply connected in the large, then there exists a universal covering $f\colon X \to Y$. A universal covering is unique up to an equivalence.*

These conditions on Y in the theorem look natural and hold for most topological spaces except "pathological" ones. From now on, we consider only topological spaces satisfying these conditions.

One of the most important features of universal coverings is their relation to fundamental groups. Let $f\colon X \to Y$ be a covering map. Fix $y_0 \in Y$ and choose an $x_0 \in f^{-1}(y_0)$. Then every path γ in Y starting at x_0 has a unique lift $\tilde{\gamma}$ in X starting at x_0 (recall that $\tilde{\gamma}$ being a lift of γ means that $f \circ \tilde{\gamma} = \gamma$). In particular, a loop with endpoints at y_0 is lifted to a curve starting at x_0 and ending at some point of $f^{-1}(y_0)$. The standard "covering homotopy" lemma tells that lifts of homotopic paths are homotopic (by a homotopy of paths we mean a homotopy with fixed endpoints). In particular, lifts of homotopic paths have the same endpoints; i.e., if two loops γ_1 and γ_2 with endpoints at y_0 are homotopic, then their lifts $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ starting at x_0 end at the same point $y \in f^{-1}(y_0)$. If f is a universal covering, then the converse statement is also true: every two paths in X connecting a given pair of points are homotopic and hence they are lifts of homotopic paths in Y.

Thus we have a 1-1 correspondence between classes of homotopic loops in Y with endpoints at y_0 (that is, the fundamental group $\pi_1(Y, y_0)$) and the elements of $f^{-1}(y_0) \subset X$.

Regular coverings and deck transformations. Actually not all coverings are equally interesting from the geometrical point of view. We want to look more closely at the class of *regular coverings*. We note in advance that all universal coverings are regular.

Recall that every continuous map $f\colon X \to Y$ induces a homomorphism of fundamental groups. Namely, if $x_0 \in X$ and $y_0 = f(x_0)$, there is a natural homomorphism $f_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0)$ which maps the class of a loop γ in X to the class of the loop $f \circ \gamma$ in Y. If f is a covering map, the covering homotopy lemma implies that the image of a noncontractible curve is noncontractible. In other words, f_* is a monomorphism (that is, an injective homomorphism) from $\pi_1(X, x_0)$ to $\pi_1(Y, y_0)$.

So the image $f_*(\pi_1(X, x_0))$ is a subgroup of $\pi_1(Y, y_0)$ isomorphic to $\pi_1(X, x_0)$. This subgroup is referred to as the *group of covering*. Geometrically, it consists of all loops with endpoints at y_0 whose lifts starting at x_0 are loops. Note that this subgroup may depend on the choice of $x_0 \in f^{-1}(y_0)$ despite the fact that all subgroups obtained this way are isomorphic.

A covering map $f\colon X \to Y$ is said to be *regular* if the following equivalent properties hold:

- $f_*(\pi_1(X, x_0))$ is a normal subgroup of $\pi_1(Y, y_0)$.
- $f_*(\pi_1(X, x_0))$ does not depend on $x_0 \in f^{-1}(y_0)$.

A *deck transformation* of a covering $f\colon X \to Y$ is a covering equivalence from X to itself, that is, a homeomorphism $h\colon X \to X$ such that $f = f \circ h$. Geometrically, a deck transformation is a homeomorphism from X to itself which "permutes the sheets" of the covering, more formally, permutes the points of $f^{-1}(y)$ for any $y \in Y$. The deck transformations obviously form a group. A covering $f\colon X \to Y$ is regular if and only if this group acts "transitively on sheets", i.e.,

- for every $y \in Y$, $x, x' \in f^{-1}(y)$ there exists a (unique) deck transformation $h\colon X \to X$ such that $h(x) = x'$.

This can be viewed as yet another equivalent definition of a regular covering.

It is not hard to show that the group of deck transformations is isomorphic to the quotient group $\pi_1(Y, y_0)/f_*(\pi_1(X, x_0))$ where $y_0 = f(x_0)$. In particular, if $f\colon X \to Y$ is a universal covering, then its group of deck transformations is isomorphic to $\pi_1(Y)$.

Example 3.4.12. For the covering $f\colon \mathbb{R} \to S^1$ given by $f(x) = (\cos x, \sin x)$ the deck transformations are maps from $\mathbb{R}$ to $\mathbb{R}$ of the form $x \mapsto x + 2\pi k$ where k is an integer. So the group of deck transformations is just $\mathbb{Z}$ acting on $\mathbb{R}$ by translations.

Similarly, the group of deck transformations of the standard covering $\mathbb{R}^2 \to T^2$ is $\mathbb{Z}^2$ acting on $\mathbb{R}^2$ by parallel translations.

For the standard covering $S^2 \to \mathbb{RP}^2$, the group of deck transformations consists of two elements, the identity and the reflection $x \mapsto -x$.

Now let Y be a length space and X be equipped with the length metric lifted from Y. Then every deck transformation is an isometry from X to itself (prove this!). In the case of a universal covering, this yields an action of $\pi_1(Y)$ on X by isometries.

This fact (that the fundamental group of a length space acts on its universal covering space by isometries) is very important. Some of its numerous applications can be found in this book.

There is another (but equivalent) approach to this subject based on the fact that the space Y and the covering $f\colon X \to Y$ are determined by the group of deck transformations. Namely Y is the quotient of X by this group action, and f is the projection to the quotient. For example, $S^1 \simeq \mathbb{R}/\mathbb{Z}$ and $T^2 \simeq \mathbb{R}^2/\mathbb{Z}^2$.

Of course, not every group action on a topological space can arise as a group of deck transformations of a covering. A necessary and sufficient condition for this is contained in the following definitions.

Definition 3.4.13. Let a group G act on a set X. This action is said to be *free* if $gx \neq x$ for all $x \in X$, $g \in G \setminus \{e\}$.

Definition 3.4.14. Let a group G act on a topological space X. This action is said to be *totally discontinuous* if every point x has a neighborhood U such that $gU \cap U = \emptyset$ for all $g \in G$ such that $gx \neq x$.

It is easy to see that a group of deck transformations of a covering acts freely and totally discontinuously. Conversely, every free totally discontinuous group action is a group of deck transformations of a covering:

Proposition 3.4.15. *Let a group G act on a topological space X freely and totally discontinuously. Then the projection $X \to X/G$ is a covering map. Moreover this is a regular covering and the group of its deck transformations coincides with G.*

Proof. The proof is trivial but it is a good test for understanding the definitions.

Let $f\colon X \to X/G$ be the projection. If $x \in X$ and U is a neighborhood of x from the definition of totally discontinuous action, then $V = f(U)$ is an evenly covered neighborhood of $f(x)$. Indeed, $f^{-1}(V)$ is the disjoint union of the sets gU over $g \in G$, and $f|_{gU}$ is a homeomorphism from gU to V.

Obviously every element of G (considered as a homeomorphism from X to itself) is a deck transformation. If $y \in Y$ and $x, x' \in f^{-1}(y)$, then there is an element $g \in G$ such that $gx = x'$. This means that G acts transitively on sheets; hence f is a regular covering. On the other hand, there is at most one deck transformation which maps x to x'; therefore every deck transformation is represented by an element of G. □

Consider a covering $f \colon X \to X/G$ where G acts on X freely and totally discontinuously. We have seen that every length metric on X/G can be lifted to X so that f becomes a local isometry. The lifted metric is *G-invariant*, which simply means that G acts by isometries with respect to this metric.

Conversely, every G-invariant length metric on X is a lift of a (unique) length metric on X/G. This is nothing but the standard metric of the quotient space defined in Section 3.3 (see also Exercise 3.4.6). We summarize these observations in the following proposition.

Proposition 3.4.16. *Let $f \colon X \to Y$ be a regular covering and G its group of deck transformations. Then the length metrics on Y are in 1-1 correspondence with the G-invariant length metrics on X so that for corresponding metrics d_X on X and d_Y on Y f is a local isometry from (X, d_X) to (Y, d_Y).*

3.4.3. Local isometries of complete spaces. In general, a local isometry is not necessarily a covering map; for instance, consider Example 3.4.3 and replace the domain of f by an open segment instead of $\mathbb{R}$. Nevertheless, for a complete length space X a local isometry $f : X \to Y$ somewhat resembles a covering map. In particular, geodesics (and shortest paths) in Y can be lifted to X:

Lemma 3.4.17. *Let $f : X \to Y$ be a local isometry of a complete length space X. Given a geodesic (shortest path) $\gamma \colon [0, a] \to Y$ and a point x_0 such that $f(x_0) = \gamma(0)$, there exists a unique geodesic (resp. shortest path) $\tilde{\gamma} : [0, a] \to X$ such that $\tilde{\gamma}(0) = x_0$ and $f(\tilde{\gamma}(t)) = \gamma(t)$.*

Proof. The proof is similar to the construction for lifts of paths in the theory of covering spaces. Assume that γ is parameterized by arc-length. Consider the subintervals $[0, t)$ of $[0, a]$ such that $\gamma_{|[0,t)}$ can be lifted to X. This means that there is a path $\tilde{\gamma}_t : [0, t) \to Y$ such that $\gamma_{|[0,t)} = f \circ \tilde{\gamma}$ together with $\tilde{\gamma}(0) = x_0$. The set of such subintervals is not empty since the restriction of f to some neighborhood of x_0 is a homeomorphism onto its image. Let $[0, t_0)$ be the union of all such subintervals, i.e., the maximum interval that admits a lifting. We have a path $\tilde{\gamma}_{t_0} : [0, t_0) \to Y$ such that $\gamma_{|[0,t)} = f \circ \tilde{\gamma}_{t_0}$. Choose a sequence $\{t_i\}$ such that $t_i < t_0$ and $t_i \to t_0$ as $i \to \infty$. Then $\tilde{\gamma}_{t_0}(t_i)$ is a Cauchy sequence and, since X is complete, it tends

to a point p. Clearly the choice of p is independent of the choice of $\{t_i\}$. Since the points $f \circ \tilde{\gamma}_{t_0}(t_i) = \gamma(t_i)$ converge to $\gamma(t_0)$ as $i \to \infty$, one can set p to be the lift of $\gamma(t_0)$ and thus γ is lifted on the closed interval $[0, t_0]$. If t_0 is not equal to a, then γ could be lifted on a larger interval $[0, t_0 + \varepsilon)$ since f is a local homeomorphism on a neighborhood of p. Since this contradicts the maximality of $[0, t_0)$, we conclude that $t_0 = a$. □

This lemma yields the following useful corollary:

Theorem 3.4.18. *Let $f : X \to Y$ be a surjective local isometry of a complete locally compact length space X. Assume that each point in Y has a neighborhood such that the shortest path between every two points from this neighborhood is unique. Then f is a covering map.*

Proof. By the definition of covering maps, we need to show that every $q \in Y$ has a neighborhood U_y whose pre-image consists of connected components each mapped homeomorphically onto U_y.

The informal idea of the proof is to represent a small ball centered at q as a bunch of shortest paths from q leading to all points of the ball. Once we choose a pre-image of q, each of the shortest paths can be lifted as a shortest path starting at this pre-image. Thus the whole ball gets lifted (as a "porcupine") to every pre-image of q.

Choose U_q to be a metric ball $B = B_r(q)$. Choose r to be so small that each point $y \in \overline{B}_r(q)$ is connected with q by a unique shortest path. Fix a point $p \in f^{-1}(q)$. By Lemma 3.4.17, for a point $y \in \overline{B}$, the shortest curve $[q, y]$ can be lifted to X as a shortest curve $[p, x]$. Introduce a map g_p given by $x = g_p(y)$ and let $V_p = g_p^{-1}(B)$ and $\overline{V}_p = g_p^{-1}(\overline{B}_r(q))$. By construction, $f(g_p(y)) = y$ and $g_p(f(x)) = x$ for $x \in \overline{V}_p$. Thus the restriction of f onto $\overline{V}_p$ is a bijection onto $\overline{B}_r(q)$ and hence a homeomorphism (since $\overline{B}_r(q)$ is compact). Thus $f_{|V_p}$ is a homeomorphism onto B.

It remains to show that all V_p are disjoint and $f^{-1}(B) = \bigcup_{p \in P} V_p$.

Indeed, assume that $x \in V_{p_1} \cap V_{p_2}$. By the construction of V_{p_1} and V_{p_2}, there are two shortest curves $[p_1, x]$ and $[p_2, x]$ that are the lifts of the shortest $[q, f(x)]$. Lifting the shortest curve $[f(x), q]$ "from the other end", that is, so that it starts at x, and using the uniqueness part of Lemma 3.4.17, we conclude that $p_1 = p_2$.

Finally, let us show that $f^{-1}(B) = \bigcup_{p \in P} V_p$. Since we know that $f(V_p) = B$ for all p, it remains to show that a point x such that $f(x) \in B$ belongs to one of the sets V_p. Consider the lift $[x, p]$ of the shortest $[f(x), q]$. Then p belongs to P and thus $x \in V_p$. □

3.5. Arcwise Isometries

Since length structure is our primary notion, we can look for maps that preserve this structure. Namely,

Definition 3.5.1. A map f between two length spaces is called an *arcwise isometry* if $L(\gamma) = L(f(\gamma))$ for every path γ.

Note that we do not require f to be bijective, since otherwise we would get nothing but usual isometries.

Exercise 3.5.2. 1. Prove that a bijective arcwise isometry is an isometry.

2. Prove that an arcwise isometry that is also a local homeomorphism is a local isometry.

3. Give a definition of local arcwise isometry and prove that local arcwise isometries are just arcwise isometries (since length structures are local).

For instance, a path parameterized by arc length is an arcwise isometry from a segment into a length space. Here we arrive at the following important notion.

3.5.1. Isometric embeddings. Injective arcwise isometries are also called *isometric embeddings.* There is often a terminological confusion around this notion. We stress that isometric embedding is *not* the same notion as *isometry onto its image.* For instance, a simple curve $\gamma : [0,1] \to \mathbb{R}^2$, $\gamma(t) = (\cos(t), \sin(t))$ is an isometric embedding, although it is not an isometry on its image (w.r.t. the restriction of Euclidean distance). Isometric embeddings that are studied in differential geometry (such as isometric embeddings of surfaces) are actually arcwise isometric embeddings. On the other hand, a sphere S^2 already does not admit an embedding $f : S^2 \to \mathbb{R}^n$ that would be an isometry onto its image. This, however, becomes possible if one substitutes an infinite-dimensional space instead of $\mathbb{R}^n$ as the target space for f. This fascinating construction is worth explaining here:

Example 3.5.3 (embedding by an isometry onto its image)**.** Let X be a compact length space. Consider the space $C(X)$ of all continuous functions $X \to R$. This space turns into a metric space by introducing the *uniform* distance between functions $d_\infty(f,g) = \sup |f(x) - g(x)|$. The map $E : X \to C(X)$ defined by $E(x) = d(x, \cdot)$ (in other notation, $E(x)$ evaluated at y is $d(x,y)$) is an embedding that is an isometry onto its image.

Exercise 3.5.4. Verify that

(i)$(C(X), d_\infty)$ is indeed a metric space. (In addition, this is also a normed vector space; we use the notation d_∞ since this is actually the L_∞ norm on this vector space).

(ii) $d_\infty(d(x,\cdot), d(y,\cdot)) = d(x,y)$, that is, $\sup_z(|d(x,z) - d(y,z)|) = d(x,y)$.

3.5.2. Surjective arcwise isometries. While bijective arcwise isometries are just isometries, surjective arcwise isometries can be extremely wild even for nice spaces. The reason for this is that they may crease the space many times. The simplest example of a pleat appears already in arcwise isometries $\mathbb{R} \to \mathbb{R}$. For instance, consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(t) = t$ if $t \le 0$, $f(t) = -t$ for $t \in [0,1]$ and $f(t) = t - 2$ if $t \ge 1$.

At first glance it may seem striking that every 2-dimensional polyhedral surface admits an arcwise isometry onto a polygonal region. This statement, however, becomes much less surprising after a practical experiment: make a model of a polyhedron using scissors, thick paper and glue, and then flatten this model by stepping on it.

It is much more surprising that every surface homeomorphic to a sphere admits an arcwise isometry onto a round sphere. This map is not at all as nice as the one that can be used for polyhedra: for a generic surface, it has everywhere dense folds and singular points. Although the proof of the existence of such arcwise isometries for surfaces is beyond the frame of this course, we sketch two absolutely elementary arguments proving that every 2-dimensional polyhedral space can be folded onto a polygon.

Proposition 3.5.5. *Every (locally-)finite 2-dimensional polyhedral space admits an arcwise isometry onto a planar polygon. In other words, every 2-dimensional polyhedral metric can be induced by a map to* $\mathbb{R}^2$.

Proof. Let P be a 2-polyhedron, and let $A = \{a_i\}$ be a finite set of its points. Define $Q_i = \{x \in P \colon |xa_i| \le |xa_j| \text{ for all } j \ne i\}$. (These sets Q_i are called Dirichlet-Voronoi regions.) It is easy to see that every set Q_i is bounded by a graph L_i whose edges are geodesic curves.

The set A can be chosen so that each Q_i possesses a triangulation with the one vertex a_i inside Q_i. To achieve this, let A be a sufficiently fine net including all vertices of P and a fine net in each edge of P. Still some pairs of points $a_i \in A$ may be connected with more than one shortest path. One can add new points to A so that, for the new set A, points $a_i, a_j \in A$ are connected with a unique shortest curve if L_i and L_j have common points.

Adding to the graph $L = \bigcup L_i$ new edges $a_i x_j$, where x_j are vertices of L_i, we obtain a triangulation of P such that every pair of triangles with different vertices a_i either have no more than one common point (vertex) or have a common edge which belongs to L.

Observe that all triangles with a common edge $[bc] \subset L$ are isometric (by the definition of Q_i). Fix a ray l in $\mathbb{R}^2$ starting at a point a. Imagine

that our polyhedron P is made of paper. Then we can cut each triangle $\triangle a_i bc$ out of P and then bend it around the bisector of the angle a_i so that directions of $a_i b$ and $a_i c$ coincide. Now we place all bent triangles in the plane $\mathbb{R}^2$ on one side of l and such that all the vertices a_i coincide with a and all the edges $[a_i b]$, $[a_i c]$ go along l. So all the edges we cut P along are placed in l and start at a. Therefore we can glue together again all such edges and this gives us the desired arcwise isometry. □

3.6. Products and Cones

3.6.1. Direct products. Among various mathematical constructions for building new objects (such as products or quotients in algebra), there are several natural ways of constructing new length spaces. The simplest of them is direct product. Let (X, d_X) and (Y, d_Y) be length spaces.

Equip the direct product $Z = X \times Y$ with the metric

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{d_X^2(x_1, x_2) + d_Y^2(y_1, y_2)},$$

where $x_1, x_2 \in X$ and $y_1, y_2 \in Y$. This formula is motivated by the Pythagorean theorem.

It is easy to see that d is a metric. This metric is called the *product* metric, and the metric space (Z, d) is called the direct metric product of X and Y.

Proposition 3.6.1. *The direct product of strictly intrinsic metrics is a strictly intrinsic metric. The direct metric product of two length spaces is a length space.*

Proof. We will carry out the argument for strictly intrinsic spaces; the second part is left as an exercise. By Theorem 2.4.16, it is sufficient to show that there is a midpoint between two given points $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ in Z. For the midpoints x_m and y_m between x_1, x_2 and y_1, y_2, resp., one sees that

$$\begin{aligned} d^2((x_m, y_m), z_1) &= d_X^2(x_m, x_1) + d_Y^2(y_m, y_1) \\ &= \frac{1}{4}(d_X^2(x_1, x_2) + d_Y^2(y_1, y_2)) = \frac{1}{4}d^2(z_1, z_2), \end{aligned}$$

and hence $d((x_m, y_m), z_1) = \frac{1}{2}(d(z_1, z_2))$. Analogously $d((x_m, y_m), z_2) = \frac{1}{2}(d(z_1, z_2))$. This shows that (x_m, y_m) is a midpoint for z_1, z_2. □

Exercise 3.6.2. To put your hand on this notion, we suggest you consider the examples of $\mathbb{R} \times S^1$, $\mathbb{R} \times S^2$, $S^2 \times S^2$. In particular, try to find all shortest curves in and all isometries of these spaces. The two latter product spaces contain a whole bunch of (nicely embedded) surfaces (tori) that are locally

isometric to the Euclidean plane. Finding these tori may be very helpful for understanding the sequel.

The product metric d constructed this way enjoys the following property: consider a fiber $S_y = \{(x, y) : y = \text{const}\}$ together with the restriction of the metric d. Then the projection map of S_y to X is an isometry. (In particular, the restriction of d is an intrinsic metric.) This follows immediately from the definition of the product metric. Evidently, the same holds for the "vertical" fibers S_x.

In addition, the group of isometries of (Z, d) is at least as rich as the product of the isometry groups of X and Y.

More precisely, every isometry I_X of X extends on Z as $I_X \times \text{id}$. Moreover, for isometries I_X of X and I_Y of Y, one canonically assigns an isometry $I_X \times I_Y : (x, y) \to (I_X(x), I_Y(y))$ of Z, and therefore the isometry group of Z contains an isomorphic copy of the product of the isometry groups of X and Y. The reader can check that, for $X = Y = \mathbb{R}$, the isometry group of the product is much richer than the product of isometry groups.

Remark 3.6.3. One should not think that our formula for product metrics is the only possible definition of a product metric on Z enjoying these nice properties. For instance, by setting $d((x, y), (x', y')) = d_X(x, x') + d_Y(y, y')$ one gets an intrinsic metric that possesses all the properties listed above. Prove this!

More generally, every norm $\| \ , \ \|$ on $\mathbb{R}^2$ gives rise to a construction of product metrics by assigning $d(z, z') = \|d_X(x, x'), d_Y(y, y')\|$. Moreover, if one axiomatizes the desirable properties of product metrics, then all possible constructions arise this way. The reason we used the Pythagorean theorem and thus the standard Euclidean norm for $\| \ , \ \|$ in our preferred definition will become clear later (as a glimpse ahead, let us mention that this is the only definition that respects the boundedness of curvature).

We leave as a not difficult exercise to justify the following description of all shortest paths of the metric space (Z, d):

Lemma 3.6.4. *A constant-speed path in Z is a shortest path (a geodesic) if and only if it is the product of two shortest paths (geodesics) in X and Y with constant-speed parameterizations.*

Notice that the projections of a shortest curve γ in Z to both X and Y are shortest paths regardless of whether the parameterization of γ is constant-speed or not. However, the other implication essentially uses this assumption: in the product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, *every* curve $(x(t), y(t))$ with increasing coordinates $x(t), y(t)$ is the product of two shortest paths $x(t)$ and $y(t)$.

Convex subsets. There is an important notion that we want to discuss in connection with the statement of the lemma. While involving this notion might seem a little bit superfluous in this simple situation, it proves very useful in more complicated cases.

We begin with the following important definition:

Definition 3.6.5. A subset A in a metric space (X, d) is said to be *convex* if the restriction of d to A is strictly intrinsic and finite.

If the metric of X is strictly intrinsic, a set A is convex if and only if for every two points $x, y \in A$, there is a shortest path between x and y which entirely belongs to A. (Prove this!) This explains the name convex.

Being convex is a "global" property of a set. Let us introduce an analogous local property.

Definition 3.6.6. Suppose that the metric of X is strictly intrinsic. A set A in X is said to be *locally convex* if every point $x \in A$ has a neighborhood U in A such that for every two points $y, z \in U$, there is a shortest path between y and z contained in A.

Exercise 3.6.7. Give a definition of local convexity for general metric spaces.

Remark 3.6.8. For a reader familiar with differential geometry, we mention that there is a special term for locally convex submanifolds (in particular, surfaces) in Riemannian and Finsler manifolds. Such submanifolds are said to be *totally geodesic* and may be alternatively defined as submanifolds in which all geodesics (of the induced metric) are geodesics of the ambient space at the same time. A submanifold of a Riemannian manifold is totally geodesic if and only if if its second fundamental form vanishes, i.e., all principal curvatures are zero.

The following lemma follows immediately from the definition:

Lemma 3.6.9. *If X is a space with strictly intrinsic finite metric and $F\colon X \to Y$ is a distance-preserving map, then the image $Im(F) := F(X)$ is convex in Y.*

Exercise 3.6.10. Prove the lemma.

Proposition 3.6.11. *Let X and Y be length spaces, and $\alpha\colon [a,b] \to X$, $\beta\colon [c,d] \to Y$ shortest paths. Then the product of their images $R = Im(\alpha) \times Im(\beta)$ is convex in $X \times Y$ and isometric to a Euclidean rectangle.*

Proof. We may assume that both α and β are parameterized by arc length. Introduce $F : [a,b] \times [c,d] \to Z$ given by $F(t,s) = (\alpha(t), \beta(s))$. This map is

an isometry:

$$d^2(F(t,s),F(t',s')) = d_X^2(g(t),g(t')) + d_Y^2(h(t),h(t')) = (t-t')^2 + (s-s')^2.$$

Applying Lemma 3.6.9 finishes the proof. □

3.6.2. Cone over a metric space. A cone $\mathrm{Con}(X)$ over a topological space X is the quotient of the product $X \times [0,\infty)$ obtained by gluing together (identifying) all points in the fiber $X \times \{0\}$. This point is called the origin (or apex) of the cone.

How should one equip a cone with a metric? To get an idea, imagine that X is a subset of the unit sphere $S^2 = \{(x,y,z) : x^2+y^2+z^2 = 1\} \subset \mathbb{R}^3$ equipped with the spherical (angular) metric. To build a cone over X, we draw a ray from the origin through every point $x \in X$. Thus a point a in the cone can be described by as (x,r), where x is a point in X that belongs to the ray Oa and $r = |aO|$ the distance from the origin; the latter is a nonnegative number. Thinking of a and x as vectors, we can write $a = tx$. How can we express the Euclidean distance between two points $a = (x,t)$ and $b = (y,s)$ in terms of x,y,t,s? Consider the triangle $\triangle Oab$. We have $|Oa| = t$, $|Ob| = s$, and the angle $\measuredangle aOb$ between the sides equals the angular distance $d(x,y)$ in X. Hence by the cosine formula

$$|ab| = \sqrt{t^2+s^2-2ts\cos(d(x,y))}.$$

We will use this formula to define cone metrics in general:

Definition 3.6.12. Let X be a metric space with $\mathrm{diam}(X) \le \pi$. The cone metric d_c on $\mathrm{Con}(X)$ is given by the formula

$$d_c(p,q) = \sqrt{t^2+s^2-2ts\cos(d(x,y))}, \tag{3.1}$$

where $p,q \in \mathrm{Con}(X)$, $p = (x,t)$, $q = (y,s)$.

Proposition 3.6.13. *If X is a metric space with* $\mathrm{diam}(X) \le \pi$, *then d_c is a metric.*

Proof. Positiveness and symmetry for d_c are trivial. Let us prove the triangle inequality. Consider three points $y_1 = (x_1,r_1)$, $y_2 = (x_2,r_2)$ and $y_3 = (x_3,r_3)$ in $\mathrm{Con}(X)$. Denote $\alpha = d(x_1,x_2)$ and $\beta = d(x_2,x_3)$. Construct three points $\bar{y}_1,\bar{y}_2,\bar{y}_3 \in \mathbb{R}^2$ so that their distances from the origin O equal r_1, r_2 and r_3 respectively, $\measuredangle\bar{y}_1O\bar{y}_2 = \alpha$, $\measuredangle\bar{y}_2O\bar{y}_3 = \beta$, and the rays $O\bar{y}_1$ and $O\bar{y}_3$ are in different half-planes with respect to $O\bar{y}_2$. Then $|\bar{y}_1\bar{y}_2| = d_c(y_1,y_2)$ and $|\bar{y}_2\bar{y}_3| = d_c(y_2,y_3)$. Now we have two cases: $\alpha+\beta \le \pi$ and $\alpha+\beta > \pi$.

If $\alpha+\beta \le \pi$, then $\measuredangle\bar{y}_1O\bar{y}_3 = \alpha+\beta \ge d(x_1,x_3)$; hence $|\bar{y}_1\bar{y}_3| \ge d_c(y_1,y_3)$. Then the triangle inequality for y_1, y_2 and y_3 follows from the triangle inequality in $\mathbb{R}^2$:

$$d_c(y_1,y_2) + d_c(y_2,y_3) = |\bar{y}_1\bar{y}_2| + |\bar{y}_2\bar{y}_3| \ge |\bar{y}_1\bar{y}_3| \ge d_c(y_1,y_3).$$

If $\alpha + \beta > \pi$, the planar triangle inequality does not help, but there is a better estimate for $|\bar{y}_1\bar{y}_2| + |\bar{y}_2\bar{y}_3|$. Indeed, since the broken line $\bar{y}_1\bar{y}_2\bar{y}_3$ lies outside the sector $\bar{y}_1 O \bar{y}_3$, we have $|\bar{y}_1\bar{y}_2| + |\bar{y}_2\bar{y}_3| \ge |\bar{y}_1 O| + |O\bar{y}_3|$. Then

$$d_c(y_1, y_2) + d_c(y_2, y_3) \ge |\bar{y}_1 O| + |O\bar{y}_3| = r_1 + r_3 \ge d_c(y_1, y_3)$$

(the last inequality follows from the definition of d_c). □

Exercise 3.6.14. Let $\tilde{\gamma} : [a, b] \to \mathrm{Con}(X)$ be a curve in the cone, $\tilde{\gamma}(t) = (\gamma(t), r(t))$ where γ is a curve in X. Prove that

$$L(\tilde{\gamma}) \ge \sqrt{r(a)^2 + r(b)^2 - 2r(a)r(b)\cos(L(\gamma))}$$

if $L(\gamma) \le \pi$, and

$$L(\tilde{\gamma}) \ge r(a) + r(b)$$

if $L(\gamma) \ge \pi$.

Hint: Repeat the argument proving the triangle inequality.

If the metric of X is intrinsic, then d_c is intrinsic too. To prove this one could write an explicit expression for a midpoint and a shortest path between two points, as we did in case of products. For cones, this is a tedious approach with cumbersome formulas. It is more convenient to begin with flat convex surfaces like those we observed in direct products.

To see where they come from, let us come back to the example of X being a subset of the unit sphere. For simplicity assume that X is the whole sphere; then the cone is the whole space $\mathbb{R}^3$. Since a shortest path γ is nothing but an arc of a greater circle, the cone over $\gamma : [0, a] \to S^2$ is a planar sector. A point in this sector has cone coordinates $(\gamma(\tau), t)$. For $\gamma(\tau)$ being parameterized by arc length, τ and t are the usual polar coordinates in the planar sector with the angular coordinate measured from the direction of the ray $[O, \gamma(0))$.

Let us see how this works in general. Let $\gamma : [0, L]$ be a shortest path in X. Introduce a polar coordinate system (r, φ) on Euclidean plane and denote by Q the set of points in the plane whose φ-coordinate is between 0 and L. Consider the map $F : Q \to \mathrm{Con}(X)$ given in polar coordinates by the formula $F(r, \varphi) = (\gamma(\varphi), r)$. The image of this map is the cone over γ.

A simple calculation shows that the map F is distance-preserving. Indeed,

$$\begin{aligned} d_c^2(F(r, \varphi), F(r', \varphi')) &= r^2 + r'^2 - 2rr' \cos d_X(\gamma(\varphi), \gamma(\varphi')) \\ &= r^2 + r'^2 - 2rr' \cos(\varphi - \varphi') = d_{\mathbb{R}^2}((r, \varphi), (r, \varphi)). \end{aligned}$$

This immediately implies that $F(Q)$ is flat and convex (Lemma 3.6.9). In particular, the image of every segment containing Q is a shortest segment in $\mathrm{Con}(X)$ and hence d_c is an intrinsic metric.

Notice that we have actually proved the following useful lemma:

Lemma 3.6.15. *If X is a length space with* $\operatorname{diam}(X) \le \pi$, *$\gamma$ is a shortest segment in X, then the cone over the image of γ is a convex flat surface in the cone $\Sigma(X)$ over X.*

Conversely, consider a shortest path $\overline{\gamma} : [a, b] \to \operatorname{Con}(X)$ in the cone not passing through the origin. We can write $\overline{\gamma}(t) = (\gamma(t), r(t))$ where $r(t) \in \mathbb{R}$ and γ is a curve in X (γ is called the *projection* of $\overline{\gamma}$ to X). Looking at the proof of the triangle inequality in the cone (Proposition 3.6.13), one sees that the triangle inequality must turn to equality for any three points $\gamma(t_1)$, $\gamma(t_2)$, $\gamma(t_3)$ such that $t_1 < t_2 < t_3$. This implies that $L(\gamma) = d(\gamma(a), \gamma(b))$; hence γ is a shortest path in X.

Thus we have a 1-1 correspondence between shortest paths in X of length strictly less than π and shortest paths in $\operatorname{Con}(X)$ not passing through the origin. As for shortest paths passing through the origin, it is easy to see the following. Every point $(x, r) \in \operatorname{Con}(X)$ is connected to the origin by a unique shortest path, namely the "segment" $\{(x, t)\}_{t \in [0,r]}$. The concatenation of two such segments with endpoints (x_1, r_1) and (x_2, r_2) is a shortest path if and only if $d(x_1, x_2) = \pi$.

Cone over a large space. Now we drop the assumption that $\operatorname{diam}(X) \le \pi$. The formula (3.1) is not suitable for defining a metric on $\operatorname{Con}(X)$ if $\operatorname{diam}(X) > \pi$; for example, the triangle inequality for d_c may fail. How does one define the cone metric in the general case? The guidelines are the following: we want the formula (3.1) to hold for "small" distances in X, and we want a cone over a length space be a length space.

If X is a length space, the existence and uniqueness of such a metric is guaranteed by Lemma 3.1.2. In fact, this metric has a simple explicit description:

Definition 3.6.16. Let (X, d) be a metric space. The cone distance $d_c(a, b)$ between points $a = (x, t)$ and $b = (y, s)$ in $\operatorname{Con}(X)$ is defined as

$$d_c(a, b) = \begin{cases} \sqrt{t^2 + s^2 - 2ts\cos(d(x, y))}, & t + s \le \pi, \\ t + s, & t + s \ge \pi. \end{cases}$$

Alternatively, one can introduce a new distance $\overline{d}$ on X by

$$\overline{d}(a, b) = \min\{d(a, b), \pi\}$$

and define $\operatorname{Con}(X, d) = \operatorname{Con}(X, \overline{d})$ where the metric on $\operatorname{Con}(X, \overline{d})$ is given by Definition 3.6.12. Since $\overline{d}$ is a metric, Proposition 3.6.13 implies that d_c is a metric. We summarize the above observations about intrinsic metrics in the following theorem.

Theorem 3.6.17. *The metric d_c on* $\mathrm{Con}(X,d)$ *is intrinsic (resp. strictly intrinsic) if and only if the metric d is intrinsic (resp. strictly intrinsic) at distances less than π. The latter means that for any $x, y \in X$ such that $d(x,y) < \pi$ there is a curve in X connecting x and y whose length is arbitrarily close (resp. is equal to) $d(x,y)$.*

Proof. First suppose that d is strictly intrinsic at distances less than π. Let $x, y \in X$, $a, b \in \mathrm{Con}(X)$, $a = (x,t)$, $b = (y,s)$. If $d(x,y) < \pi$, apply Lemma 3.6.15 to a shortest path γ connecting x and y. It follows that there is a curve of length $d_c(a,b)$ connecting a and b. If $d(x,y) \geq \pi$, then $d_c(a,b) = t+s$ and there is a curve of length $t+s$ connecting a and b, namely, the union of the two segments connecting a and b to the origin. Thus d_c is strictly intrinsic.

Conversely, suppose that the metric d_c is strictly intrinsic. For any two points $x, y \in X$ with $d(x,y) < \pi$ apply the result of Exercise 3.6.14 to a shortest path $\tilde{\gamma}$ connecting the points $a = (x,1)$ and $b = (y,1)$ in the cone. Since $L(\gamma) = d_c(a,b) < 2$, $\tilde{\gamma}$ does not pass through the origin and hence has a well-defined (and continuous) projection γ in X. The inequality from Exercise 3.6.14 implies that $L(\gamma) = d(x,y)$.

The proof for non-strictly intrinsic metrics is similar and is left to the reader. □

Exercise 3.6.18. Let X be a metric space with $\mathrm{diam}(X) = \pi$. Suppose that $\mathrm{Con}(X)$ is a length space but X is not. Prove that there are three distinct points $x, y, z \in X$ such that $|xy| = |xz| = \pi$.

Exercise 3.6.19. Let X be a line segment of length α, $0 < \alpha < 2\pi$. Prove that $\mathrm{Con}(X)$ is isometric to the planar sector of angular measure α with its intrinsic metric.

This exercise is a partial case of the next proposition which says that every polyhedral space locally looks like a cone. For simplicity we restrict ourselves to 2-dimensional polyhedral spaces. In higher dimensions, one should consider cones over *spherical* polyhedral spaces.

Proposition 3.6.20. *Let $p \in P$ be a point in a two-dimensional polyhedral space P. For all sufficiently small $r > 0$, the ball $B_r(p)$ is isometric to an r-ball in* $\mathrm{Con}(G)$ *centered at the origin of the cone over a graph G.*

Exercise 3.6.21. Prove the proposition.

Hint: First prove that the construction of a cone "commutes" with gluing; i.e., if a space space X is obtained by gluing from a space Y, then the cone $\mathrm{Con}(X)$ can be obtained by gluing respective rays in the cone $\mathrm{Con}(Y)$.

The graph G mentioned in the proposition is called the *link* of P at p. If p belongs to the interior of a face, the link is just a circle of length 2π. For a

point in the interior of an edge, the link is a graph consisting of two vertices connected by several segments of length π (each segment corresponds to a face of P adjacent to the edge containing p). If p is a vertex, any graph may appear as the link at p.

3.6.3. Spherical suspensions. There are other constructions with familiar Euclidean ancestors. For instance, let us simulate the procedure building S^n out of its equator S^{n-1} by means of adding two poles and drawing a semicircle (meridian) through every point of S^{n-1}. Begin with the direct product $X \times I$ of a topological space X and a segment $I = [0, a]$ and contract the fibers $X \times 0$ and $X \times a$ each to a point. The resulting space $\Sigma(X)$ is called the *spherical cone*, or the *suspension*, over X.

If (X, d) is a length space with $\operatorname{diam} X \le \pi$, we choose $a = \pi$ and define the distance d_Σ by the equation

$$\cos d_\Sigma(p, q) = \cos t \cos s + \sin t \sin s \cos d(x, y)$$

for all points $p = (x, t)$, $q = (y, s)$ in $\Sigma(X)$. This formula is certainly suggested by the cosine theorem in *spherical* geometry.

Exercise 3.6.22. 1. For the standard sphere S^n, prove that $\Sigma(S^n)$ is isometric to the standard sphere S^{n+1}.

2. Similarly as for cones and products, the suspension over a geodesic in X is a convex (totally geodesic) surface in $\Sigma(X)$. In this case, however, it is not flat. This surface is locally isometric to the standard unit sphere. Prove these statements.

3. Arguing as in the cone case, prove that d_Σ is an intrinsic metric.

Notice: This exercise assumes some knowledge of spherical geometry, such as the spherical cosine formula.

3.6.4. Warped products. Let us mention a construction which generalizes direct (metric) products, cones, and suspensions.

Let X and Y be two (complete) length spaces and $f\colon X \to \mathbb{R}$ a positive continuous function. Consider a Lipschitz curve γ in $X \times Y$ whose projections to X and Y are $\gamma_1\colon [a, b] \to X$ and $\gamma_2\colon [a, b] \to Y$. To equip $X \times Y$ with a warped product metric, define the length of γ by the formula

$$L(\gamma) = \int_a^b \sqrt{|\gamma_1'|^2(t) + f^2(\gamma_1(t))|\gamma_2'|^2(t)}\, dt,$$

where $|\gamma_1'|$, $|\gamma_2'|$ are defined a.e. (see 2.7.6). The metric on $X \times Y$ induced by this length structure is called the *warped product metric*. One denotes by $M \times_f N$ the space $X \times Y$ equipped with this metric.

It is obvious that direct metric product is a particular case of warped product (with $f \equiv 1$).

Instead of a positive function f, it is sometimes natural to consider f vanishing at certain points. Then one obtains $X \times Y$ with a warped product semi-metric. As usual, this semi-metric in its turn gives rise to a metric after points at zero distance from each other are identified. This new space with the resulting metric will also be called a warped product.

Exercise 3.6.23. Let $\mathrm{Con}(X)$ be the cone over a length space X, and $\operatorname{diam} X < \pi$. Prove that $\mathrm{Con}(X) = [0, \infty) \times_f X$, where $f(t) = t$.

Exercise 3.6.24. Let $\Sigma(X)$ be the spherical suspension over a length space X. Show that $\Sigma(X) = [0, \pi] \times_f X$, where $f(t) = \sin t$.

3.6.5. Definition of angle. Since the notion of angle belongs to basic notions of Euclidean geometry, it is desirable to define angles in our metric context. To figure out how we could measure angles in a metric space, let us first express usual Euclidean angles in purely metric terms. Consider two rays $\alpha : [0, \infty[\to \mathbb{R}^2$ and $\beta : [0, \infty[\to \mathbb{R}^2$ emanating from the same point $a = \alpha(0) = \beta(0)$. Picking two positive numbers t, s and applying the cosine theorem to the triangle $a\alpha(t)\beta(s)$, we express the angle between the rays as

$$\arccos \frac{|a\alpha(t)|^2 + |a\beta(s)|^2 - |\alpha(t)\beta(s)|^2}{2|a\alpha(t)||a\beta(s)|}.$$

The fact that this expressions is independent of the choice of t and s is certainly a lucky coincidence: should we use two smooth curves instead of rays, we would get a nonconstant function in t and s; thinking of angle as an infinitesimal notion suggests passing to the limit as t and s tend to zero.

Definition 3.6.25. Let x, y, z be three distinct points in a metric space (X, d). The *comparison angle* xyz, denoted by $\tilde{\angle} xyz$ or $\tilde{\angle}(x, y, z)$, is defined by

$$\tilde{\angle} xyz = \arccos \frac{d(x,y)^2 + d(y,z)^2 - d(x,z)^2}{2\, d(x,y)\, d(y,z)}.$$

The geometric meaning of this definition is the following. Let $\triangle \bar{x}\bar{y}\bar{z}$ be a triangle in $\mathbb{R}^2$ whose sides $|\bar{x}\bar{y}|$, $|\bar{y}\bar{z}|$ and $|\bar{x}\bar{z}|$ equal the respective distances $d(x, y)$, $d(y, z)$ and $d(x, z)$. (Such a triangle is uniquely defined up to a rigid motion.) Then $\tilde{\angle} xyz = \angle \bar{x}\bar{y}\bar{z}$.

Definition 3.6.26. Let $\alpha : [0, \varepsilon) \to X$ and $\beta : [0, \varepsilon) \to X$ be two paths in a length space X emanating from the same point $p = \alpha(0) = \beta(0)$. We define the angle $\angle(\alpha, \beta)$ between α and β as

$$\angle(\alpha, \beta) = \lim_{s,t \to 0} \tilde{\angle}(\alpha(s), p, \beta(t))$$

if the limit exists.

Let us introduce some notation. Assuming the curves α and β fixed, define

$$\theta(s,t) = \widetilde{\measuredangle}(\alpha(s), p, \beta(t)).$$

In this notation,

$$\measuredangle(\alpha,\beta) = \lim_{s,t\to 0, s\to 0} \theta(s,t).$$

If α and β are shortest paths parameterized by arc length, then $d(p,\alpha(s)) = s$, $d(p,\beta(t)) = t$, and $\theta(s,t)$ is determined by the distance $d(\alpha(s),\beta(t))$:

$$\theta(s,t) = \arccos \frac{s^2 + t^2 - d(\alpha(s),\beta(t))^2}{2st}.$$

Then the definition of angle (for unit-speed curves) can be rewritten as follows: the angle $\measuredangle(\alpha,\beta)$ exists and equals $\theta_0 \in [0,\pi]$ if and only if

$$d(\alpha(s),\beta(t))^2 = s^2 + t^2 - 2st\cos\theta_0 + o(st), \qquad s,t \to 0.$$

The notion of angle will be mainly used for shortest paths, but let us first analyze what we get for different paths in usual Euclidean space. It turns out that the existence of angle is closely connected with differentiability:

Proposition 3.6.27. *Let $\alpha\colon [0,\varepsilon) \to \mathbb{R}^2$ and $\beta : [0,\varepsilon) \to \mathbb{R}^2$ be two paths parameterized by arc length and emanating from the same point $a = \alpha(0) = \beta(0)$. Then*

1. If both paths are differentiable at $t = 0$, then the angle $\measuredangle(\alpha,\beta)$ is equal to the angle between their velocity vectors.

2. If at least one of the paths α, β is not differentiable at $t = 0$, then the angle $\measuredangle(\alpha,\beta)$ does not exist.

Exercise 3.6.28. Prove the proposition.

There are lots of examples of length spaces where the angle $\measuredangle(\alpha,\beta)$ does not exist even for two shortest curves α,β. Perhaps the simplest example of this type can be made out of two copies of $\mathbb{R}$ by gluing them together at the origin and at all points of the form 2^{-n} with integer n.

Another example is given in the following exercise:

Exercise 3.6.29. Let $(V,|\cdot|)$ be a two-dimensional normed space. Prove that either $(V,|\cdot|)$ is a Euclidean space or it contains two rays such that the angle between them is not defined.

Hint: A normed space is Euclidean if and only if

$$|w+v|^2 + |v-w|^2 = 2(|v|^2 + |w|^2)$$

for all vectors v, w.

The following natural properties of angles follow immediately from the definition and are left as easy exercises:

Proposition 3.6.30. *1. Every shortest path forms zero angle with itself.*

2. If two shortest segments $[a,b]$ and $[b,c]$ are such that their concatenation ("$[abc]$") is also a shortest path, then the angle between $[b,a]$ and $[b,c]$ is π.

Remark 3.6.31. In this course we will mainly deal with length spaces with certain curvature bounds. In such spaces the angle between two shortest paths is always well-defined. For general spaces one may consider *upper* angles, which always exist. Namely,

Definition 3.6.32. The *upper angle* $\measuredangle_U(\alpha,\beta)$ is defined as

$$\measuredangle_U(\alpha,\beta) = \limsup_{s,t\to 0} \widetilde{\measuredangle}(\alpha(s),p,\beta(t)).$$

Remark 3.6.33. Historically the notion of angle played a very important role in the development of metric geometry. It is now well understood that the usage of angles can be eliminated (without essential complications) from the most arguments dealing with singular (non-Riemannian) spaces. On the other hand, involving angles often helps to visualize arguments and evokes Riemannian analogies.

3.6.6. Space of directions. As we mentioned above, the notion of angle does not play a very important role in modern metric geometry. Nevertheless, we will use it to define the cornerstone notion of the space of directions. In a sense, this notion replaces the concept of tangent space in the theory of smooth manifolds, and the advanced theory of Alexandrov spaces begins with this notion.

Let us fix a point p and consider the space of all curves starting from p. Our goal is to choose a subspace of "nicely behaved curves" and put an angular semi-metric on this space.

To do this we need a version of the triangle inequality for angles.

Important remark: Please pay very close attention to the proof of this theorem. This is the first example of an argument based on comparing certain configurations of points in a length space with in some sense analogous configurations in the Euclidean plane. This type of arguments is as basic and widely used in metric geometry as ε-δ arguments in analysis.

Theorem 3.6.34. *Consider three curves γ_1,γ_2 and γ_3 starting at p. Assume that the angles $\alpha_1 = \measuredangle(\gamma_2,\gamma_3)$, $\alpha_2 = \measuredangle(\gamma_1,\gamma_3)$ exist. If the angle $\alpha_3 = \measuredangle(\gamma_1,\gamma_2)$ also exists, then it satisfies the following triangle inequality:*

$$\alpha_3 \le \alpha_1 + \alpha_2.$$

Remark 3.6.35. One of the assumptions of this theorem is the existence of the angles α_1 and α_2. In fact, the same inequality holds for upper angles, which always exist.

Proof. The statement is trivial if $\alpha_1 + \alpha_2 \geq \pi$. So suppose that this is not the case.

Each angle is the limit of an appropriate function θ. Thus, given a positive ε, for all sufficiently small s, t, r one has

$$|\alpha_1 - \theta(b,c)| \leq \varepsilon, |\alpha_2 - \theta(a,c)| \leq \varepsilon, |\alpha_3 - \theta(a,b)| \leq \varepsilon,$$

where $a = a(s) = \gamma_1(s)$, $b = b(t) = \gamma_2(t)$ and $c = c(r) = \gamma_3(r)$.

Here the comparison part begins! Let us develop two "triangles" $\triangle pac$ and $\triangle pbc$ on the Euclidean plane. This means that we pick four points $\bar{p}$, $\bar{a}$, $\bar{b}$ and $\bar{c}$ in the Euclidean plane $\mathbb{R}^2$ so that

$$|\bar{p}\bar{a}| = d(p,a), \ |\bar{p}\bar{b}| = d(p,b), \ |\bar{p}\bar{c}| = d(p,c), \ |\bar{c}\bar{a}| = d(c,a), \ |\bar{c}\bar{b}| = d(c,b)$$

and the points $\bar{a}$ and $\bar{b}$ are situated on opposite sides of the line $(\bar{p}\bar{c})$.

Let us fix a, b and move c towards p. Formally this means that we fix s and t and decrease r. For c very close to p (while a and b are fixed) it is easy to see that $\bar{p}$ and $\bar{c}$ are situated on the same side of the line $\bar{a}\bar{b}$. (Drawing a picture here is a must!)

On the other hand, fixing r and making s and t sufficiently small we obtain a configuration with $\bar{p}$ and $\bar{c}$ lying on opposite sides of the line $\bar{a}\bar{b}$.

By continuity, we can find such values of the parameters s, t, r that $\bar{c}$ belongs to the segment $[\bar{a}, \bar{b}]$. To formalize this, we can fix the choice of $\bar{p}$; for instance choose $\bar{p} = (0,0) \in \mathbb{R}^2$ and choose $\bar{c}$ in a fixed ray, say $\bar{c} = (d(p,c), 0)$. Then the choices of $\bar{a}$ and $\bar{b}$ are uniquely determined by and continuously depend on the distances $d(p,a), d(p,b), d(c,a), d(c,b)$ and $d(p,c)$. Thus all four points move continuously (well, $\bar{p}$ does not move at all) as we vary s, t, r.

For $\bar{c} \in [\bar{a}, \bar{b}]$, we obviously have $|\bar{a}\bar{c}| + |\bar{c}\bar{b}| = |\bar{a}\bar{b}|$, and hence

$$|\bar{a}\bar{b}| = |\bar{a}\bar{c}| + |\bar{c}\bar{b}| = d(a,c) + d(c,b) \geq d(a,b).$$

Add to our planar configuration a point $\tilde{b}$ such that

$$|\bar{p}\tilde{b}| = |\bar{p}\bar{b}| = d(p,b), \quad |\bar{a}\tilde{b}| = d(a,b)$$

and $\tilde{b}$ lies on the same side of the line $(\bar{p}\bar{a})$ as $\bar{b}$.

Recall that $\theta(a,b)$ is by definition equal to the angle $\measuredangle \bar{a}\bar{p}\tilde{b}$ of the triangle $\triangle \tilde{b}\bar{p}\bar{a}$ at $\bar{p}$, and analogously $\theta(a,c) = \measuredangle \bar{a}\bar{p}\bar{c}$, $\theta(b,c) = \measuredangle \bar{b}\bar{p}\bar{c}$. Therefore

$$\theta(a,c) + \theta(b,c) = \measuredangle \bar{a}\bar{p}\bar{b}.$$

Comparing the triangles $\triangle \bar{b}\bar{p}\bar{a}$ and $\triangle \tilde{b}\bar{p}\bar{a}$, we see that they have two equal sides and $|\bar{a}\bar{b}| \geq |\bar{a}\tilde{b}|$. Hence their angles also satisfy the inequality $\measuredangle \bar{a}\bar{p}\bar{b} \geq \measuredangle \bar{a}\bar{p}\tilde{b}$. Thus we obtain that

$$\theta(a,c) + \theta(b,c) \geq \theta(a,c).$$

Combining this with

$$|\alpha_1 - \theta(b,c)| \le \varepsilon, |\alpha_2 - \theta(a,c)| \le \varepsilon, |\alpha_3 - \theta(a,b)| \le \varepsilon,$$

we see that $\alpha_3 \le \alpha_1 + \alpha_2 + 3\varepsilon$ for all positive ε. This completes the proof of the theorem. □

Exercise 3.6.36. Prove the following proposition strengthening the theorem:

Proposition 3.6.37. *1. Theorem 3.6.34 holds for upper angles.*

2. Under the assumptions of Theorem 3.6.34, suppose that at least one of the angles α_1 and α_2 is equal to zero. Then α_3 also exists.

Exercise 3.6.38. Prove that the sum of adjacent angles is at least π. More precisely, if two shortest segments $[a,b]$ and $[b,c]$ are such that their concatenation ("$[abc]$") is also a shortest path, and the angles $\measuredangle dba$ and $\measuredangle dbc$ exist, then their sum is greater than or equal to π.

Definition 3.6.39. We say that a curve γ (starting at p) *has a direction* at p if the angle $\measuredangle(\gamma,\gamma)$ does exist. We say that two curves α, β have the same direction at p if the angle $\measuredangle(\alpha,\beta)$ exists and is equal to 0.

Exercise 3.6.40. (i) Give an example of a curve $\gamma : [0,\varepsilon) \to \mathbb{R}^2$ such that the angle $\measuredangle(\gamma,\gamma)$ (at $\gamma(0)$) does not exist.

(ii) Prove that $\measuredangle(\gamma,\gamma)$, if it exists, is equal to zero.

(iii) Prove that if γ is a geodesic, then the angle $\measuredangle(\gamma,\gamma)$ always exists.

Now we consider only curves having direction and observe that the property of having the same direction is an equivalence relation. A class of equivalent curves is called a *direction*. One notices that this space of directions is a metric space with respect to the upper angle.

Unfortunately, this seemingly very general notion has not proven useful (as perhaps it is indeed *too* general). As a glimpse ahead, let us informally explain what we will use instead of this space. We will mainly study such spaces where the angle between shortest paths always exists. We will restrict ourselves to such curves that are equivalent to a shortest curve. In other words, we care only for directions arising from shortest paths. The space of such directions will be a metric space with respect to angle (vs. upper-angle in the general case). Alas, this is not yet the space of directions, as this space may be noncomplete, and only its completion is called *the space of directions* or directional space.

Chapter 4

Spaces of Bounded Curvature

4.1. Definitions

4.1.1. Introduction. General length spaces can be extremely nasty, and there are almost no nontrivial results which do not impose additional restrictions on spaces in question. Among such geometric restrictions, which serve as sort of regularity assumptions, the main role is played by *curvature bounds*. Loosely speaking, curvature bounds guarantee a certain degree of convexity or concavity for distance functions. We begin with the two most important curvature bounds: spaces of nonpositive curvature and spaces of nonnegative curvature. These are two classes of spaces which enjoy very different and distinct features, and whose analysis however involves very similar machinery. While other classes of curvature bounds are also very important, most of their main features can be traced in the two main classes. Spaces with a curvature bound (either above or below) will be called *Alexandrov spaces*.

At the first glance, spaces of bounded curvature may have neither topologic nor metric resemblance with Euclidean spaces. Nevertheless, their definition is based on comparisons with the Euclidean world and (in certain respects) they are less monstrous than, for instance, normed spaces. For instance, the notion of angle makes perfect sense in spaces of bounded curvature, while there are very serious reasons why this notion cannot be generalized to normed spaces.

We give several equivalent definitions of nonpositively (resp. nonnegatively) curved spaces. These definitions formalize that:

- Distance functions for such spaces are not less convex (resp. not less concave) than for the Euclidean plane;
- Geodesics emanating from one point diverge at least as fast as (resp. not faster) than in the Euclidean plane;
- Triangles are not thicker (resp. not thinner) than Euclidean triangles with the same side lengths.

Very vaguely, these spaces are not smaller (resp. not larger) than the Euclidean space: for instance, they admit more nonexpanding (resp. expanding) maps.

The sphere is an example of a positively curved space where the above properties are seen easily. Spherical triangles look "fat" compared to their Euclidean counterparts (for example, the angles in spherical triangles are greater than those in Euclidean ones). Spherical geodesics tend to "attract" one another rather than diverge linearly (more precisely, the distance between two geodesics emanating from one point in the sphere is a concave function). These are typical features of spaces having positive curvature.

More generally, every convex surface in $\mathbb{R}^3$ (that is, a boundary of a convex body) is a nonnegatively curved space, and every smooth saddle surface (locally looking like a hyperbolic paraboloid) is a nonpositively curved space. For a reader with a background in Riemannian geometry, we mention in advance that a Riemannian manifold is a nonnegatively (resp. nonpositively) curved length space if and only if its sectional curvatures are nonnegative (resp. nonpositive).

Convention. Unless otherwise stated, all length spaces in this chapter are assumed connected (or, equivalently, with finite metrics), and their metrics are assumed strictly intrinsic, i.e., such that any two points can be connected by a shortest path.

With this convention we sacrifice generality for the sake of simplicity of our exposition; more general cases will be considered later (namely in Chapters 9 and 10).

4.1.2. Comparisons for distance functions. Fix a point $p \in X$; call p a reference point. Distance to p is a real-valued function defined on X: $d_p(x) = d(x,p)$. It is easier to deal with functions $\mathbb{R} \to \mathbb{R}$, and to study d_p restricted to various shortest segments in X. More precisely, for a shortest path $[ab] = \gamma \colon [0,T] \to X$ parameterized by arc length, we introduce a function $g(t) = d(p,\gamma(t)) = d_p \circ \gamma(t)$, which represents the distance function d_p restricted to the segment γ. We will call such functions *1-dimensional distance functions.*

We want to compare g with an appropriate 1-dimensional Euclidean distance function. To do so, we choose a segment of the same length in the Euclidean plane and a reference point "positioned in the same way as p is positioned with respect to γ". Formally, we choose a Euclidean *comparison* segment $[\bar{a}\bar{b}]$ of the same length T as $[a,b]$ and a reference point $\bar{p}$ such that $|\bar{a}\bar{p}| = d_p(a) = d(a,p)$ and $|\bar{b}\bar{p}| = d_p(b) = d(b,p)$. Of course, this comparison configuration is unique up to a rigid motion. Regard this segment as a path $\gamma_0(t)$ parameterized by arc length; $\gamma_0(0) = \bar{a}$, $\gamma_0(T) = \bar{b}$.

Definition 4.1.1. The *comparison function* for g is $g_0(t) = |\bar{p}\gamma_0(t)|$, the Euclidean distance from $\bar{p}$ restricted to a comparison segment $[\bar{a}\bar{b}]$.

Following our ideology that distance functions for nonpositively (resp. nonnegatively) curved spaces must be more convex (resp. more concave) than for the Euclidean plane, we are going to define such spaces by the inequality $g_0(t) \le g(t)$ (resp. $g_0(t) \ge g(t)$). In order to make our definition local, we do not impose this condition on all pairs (p,γ); we require this to hold only in a sufficiently small neighborhood of each point.

Definition 4.1.2 ("*Distance condition*")**.** We say that (X,d) is *nonpositively* (resp. *nonnegatively*) *curved* if every point in X has a neighborhood such that, whenever a point p and a segment γ lie within this neighborhood, the comparison function g_0 for the 1-dimensional distance function $g(t) = d(p,\gamma(t))$ satisfies $g_0(t) \le g(t)$ (resp. $g_0(t) \ge g(t)$) for all $t \in [0,T]$.

We will use the name *Alexandrov space* for all spaces with a curvature bound, and in particular for spaces of nonpositive and nonnegative curvature.

4.1.3. First examples. Though we choose very simple examples to begin with, the proofs of their properties are rather long (especially for the second one). If your intuition tells you that these examples are correct, you can postpone proofs for a while. The reason is that in the proofs we use the definition straightforwardly, working by "bare hands", not having yet any machinery to shorten them.

Example 4.1.3. Attach together three copies of the ray $[0,\infty) \subset \mathbb{R}$ by gluing at the point 0. The resulting space $R_{(3)}$ has nonpositive curvature.

Example 4.1.4. Let K be the cone over a circle of length L (see 3.6.2). Then K is a space of nonnegative curvature if $L \le 2\pi$ and K is a space of nonpositive curvature if $L \ge 2\pi$.

Example 4.1.5. Consider $\mathbb{R}^2$ with the norm $\|v\| = |x| + |y|$ where x and y are Cartesian coordinates of v. This normed space X is neither nonnegatively curved nor nonpositively curved.

Now we pass to the proofs for these examples.

Proof for Example 4.1.3. Denote by O the common point of the three rays. Every shortest path in $R_{(3)}$ is either a segment in one of the rays, or a concatenation of two segments in two different rays. Let $\gamma\colon [0,T] \to R_{(3)}$ be a shortest path and $p \in R_{(3)}$. If two of the points $\gamma(0)$, $\gamma(T)$ and p belong to the same ray, then the statement is trivial because γ and p are contained in a union of two rays and this union is isometric to $\mathbb{R}$. So we consider only the case when the three points p, $a = \gamma(0)$, and $b = \gamma(T)$ belong to different rays. For every $x \in [Oa]$ one has $|px| = |pa| - |ax|$. For the function g from Definition 4.1.2, this means that $g(t) = g(0) - t$ if $\gamma(t) \in [Oa]$. For the function g_0, one has $g_0(t) \geq g_0(0) - t$ by the triangle inequality. Since $g_0(0) = g(0)$, the desired inequality $g(t) \leq g_0(t)$ follows for $\gamma(t) \in [Oa]$. The case $\gamma(t) \in [Ob]$ is similar. □

Proof for Example 4.1.4. The cone over a circle is flat outside the vertex; every sub-cone over a segment of length $\alpha \leq \max\{L/2, \pi\}$ is convex and isometric to a planar sector of angular measure α. (Recall the discussion of cones in Section 3.6.2.) For a shortest path $\gamma\colon [0,T] \to K$ and a point $p \in K$, consider a triangle $\triangle$ composed of three shortest paths between the points p, $a = \gamma(0)$ and $b = \gamma(T)$. There are two possibilities:

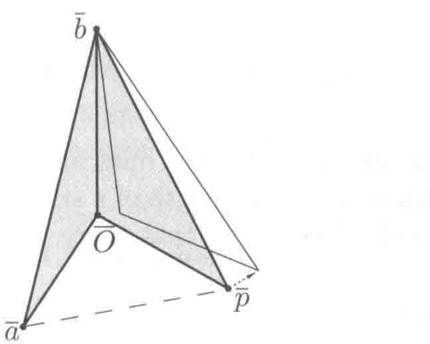
- The triangle $\triangle$ bounds a region not containing O, or one of the points a, b, p coincides with O.
- The triangle $\triangle$ bounds a region containing O, or some of its sides pass through O.

In the first case, the bounded region is isometric to (the region bounded by) a triangle in the plane, and functions g and g_0 from Definition 4.1.2 coincide.

We consider the second case separately for $L < 2\pi$ and $L > 2\pi$. (The case $L = 2\pi$ is trivial because the cone is isometric to $\mathbb{R}^2$.)

1. Suppose that $L < 2\pi$. Then one can represent the cone K as a trihedral cone in R^3 with edges passing through the vertices of $\triangle$. (The three-hedral cone is regarded with its induced length metric; "represent" means that the two cones are isometric.) To do this, cut K into three sectors by the rays Oa, Ob and Op, and place these sectors in $\mathbb{R}^3$ so as to form a three-hedral angle (this is possible due to the triangle inequality for angles).

Now the sides of $\triangle$ are straight segments in $\mathbb{R}^3$. These three segments are contained in some plane $P \subset \mathbb{R}^3$, and one can use the same triangle, regarded as a subset of P, to define the function g_0 from Definition 4.1.2. In other words, $g_0(t)$ equals the distance in $\mathbb{R}^3$ from p to $\gamma(t)$. Since the



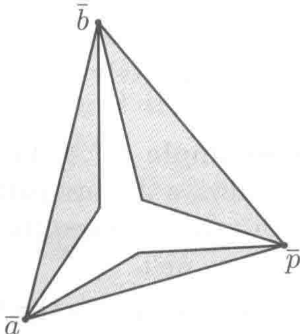


Figure 4.1: Three comparison triangles do not overlap.

distances in the cone's intrinsic metric are greater than or equal to those in the ambient space $\mathbb{R}^3$, it follows that $g \geq g_0$. Therefore the cone is a space of nonnegative curvature.

2. Suppose that $L > 2\pi$. The triangles $\triangle abO$, $\triangle apO$ and $\triangle bpO$ are flat, i.e., isometric to planar ones. Consider the triangles $\triangle abO$ and $\triangle bpO$ and place their isometric copies $\triangle\bar{a}\bar{b}\bar{O}$ and $\triangle\bar{b}\bar{p}\bar{O}$ in the plane at different sides of the common side $\bar{O}\bar{b}$. (If a shortest path $[ab]$ or $[bp]$ passes through O, its isometric copy degenerates to a segment.) Observe that $\measuredangle\bar{a}\bar{O}\bar{b} + \measuredangle\bar{b}\bar{O}\bar{p} > \pi$ and $\measuredangle\bar{a}\bar{O}\bar{p} \leq \measuredangle aOc$; hence $|\bar{a}\bar{p}| \leq |ap|$. Let us rotate the triangle $\triangle\bar{a}\bar{b}\bar{O}$ around $\bar{b}$ until $|\bar{a}\bar{p}|$ becomes equal to $|ac|$ (see the left part of Figure 4.1). This shows that isometric copies of $\triangle abO$ and $\triangle bpO$ lie without overlapping in a planar triangle $\triangle\bar{a}\bar{b}\bar{p}$ whose sides are equal to those of $\triangle abp$. This arguments work for any pair of triangles $\triangle aOb$, $\triangle bOp$, $\triangle pOa$; hence their isometric copies lie in $\triangle\bar{a}\bar{b}\bar{c}$ without overlapping, as shown in the right part of Figure 4.1.

Then it is clear that all distances between points in the sides of $\triangle abp$ are less than or equal to distances between corresponding points of the comparison triangle $\triangle\bar{a}\bar{b}\bar{p}$, i.e., $g \leq g_0$. Therefore the cone is a space of nonpositive curvature. $\square$

Remark 4.1.6. From the above proof one can see that the converse statements are also true: if a cone over a circle is nonnegatively (resp. nonpositively) curved, then the length of the circle is not greater (resp. not less) than 2π.

The following exercise tells one about curvature of a cone over a segment (note that a cone over a sufficiently short segment is just a planar sector). We will use the result of this exercise later in this chapter.

Exercise 4.1.7. Let X be a cone over a segment of length L. Prove that

1. X is nonpositively curved for any L.
2. X is nonnegatively curved if and only if $L \leq \pi$.

Proof for Example 4.1.5. First note that straight lines in a normed vector space are always shortest paths, because the length of a straight segment equals the distance between its endpoints. Note that there may be other shortest paths as well.

The unit sphere in X is the Euclidean square with vertices $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$. Let $p = (0,0)$. First consider the shortest path connecting the points $(1,0)$ and $(0,1)$, parameterized by the interval $[-1,1]$. For this shortest path, $g(t) \equiv 1$ while $g_0(t) = |t|$, so $g > g_0$ except endpoints, contrary to the definition of nonpositive curvature. Then consider the shortest path connecting points $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{2}, -\frac{1}{2})$ and parameterized by the interval $[0,1]$. Here $g(0) = g_0(0) = 1$ and $g(1) = g_0(1) = 1$. However $g(\frac{1}{2}) = \frac{1}{2}$ while $g_0(\frac{1}{2}) = \frac{\sqrt{3}}{2} < \frac{1}{2}$. So in this case $g(\frac{1}{2}) < g_0(\frac{1}{2})$, contrary to the definition of nonnegative curvature.

The definitions of nonpositive and nonnegative curvature are local, but one readily sees that the same construction can be repeated in an arbitrary neighborhood of the origin since Euclidean homotheties are homotheties w.r.t. the norm as well. □

4.1.4. Distance comparison for triangles. Probably the reader has not yet gained much insight into the geometry of nonpositively (nonnegatively) curved spaces from our analysis of distance functions. Let us re-formulate our definition in more geometric terms. Traditionally the assertion of these definitions is abbreviated as the CAT(0)-condition, and spaces fitting these definitions are called CAT(0)-spaces. Here CAT stands for the comparison of Cartan–Alexandrov–Toponogov, and (0) means that we impose zero as the curvature bound, reflecting the fact that we compare our space with the Euclidean plane. Comparing with other spaces (such as spheres) one defines other CAT(k)-spaces, $k \in \mathbb{R}$. This abbreviation is usually used for upper curvature bounds only, whereas in case of lower curvature bounds one just speaks of Alexandrov spaces of curvature bounded below by k. Note that the term "Alexandrov space" can often be confusing since it does not indicate whether k is the upper bound or the lower bound, and this must be specified.

By a triangle in X we mean a collection of three points, a, b and c (vertices) connected by three shortest paths (sides). For brevity we denote these shortest segments by $[ab]$, $[bc]$, $[ca]$ and their lengths by $|ab|$, $|bc|$, $|ca|$, respectively. Recall that the vertices alone may not define a triangle uniquely since there may be several different shortest paths between the same pair of

vertices. By $\measuredangle abc$ we denote the angle between the shortest paths $[ba]$ and $[bc]$ at b (if this angle is well-defined).

For each triangle $\triangle abc$ in X, we construct a triangle $\triangle \bar{a}\bar{b}\bar{c}$ in the Euclidean plane with the same lengths of sides, i.e.,

$$|ab| = |\bar{a}\bar{b}|, \quad |bc| = |\bar{b}\bar{c}|, \quad |ac| = |\bar{a}\bar{c}|.$$

Definition 4.1.8. Such a triangle $\triangle \bar{a}\bar{b}\bar{c}$ is called a *comparison* triangle for the triangle $\triangle abc$.

It is clear that a comparison triangle is uniquely defined up to a rigid motion of Euclidean plane. Now we can re-formulate the comparison condition for distance functions as follows.

Definition 4.1.9 ("*Triangle condition*"). X is a space of nonpositive (resp. nonnegative) curvature if in some neighborhood of each point the following holds:

For every $\triangle abc$ and every point $d \in [ac]$, one has $|db| \le |\bar{d}\bar{b}|$ (resp. $|db| \ge |\bar{d}\bar{b}|$) where $\bar{d}$ is the point on the side $[\bar{a}\bar{c}]$ of a comparison triangle $\triangle \bar{a}\bar{b}\bar{c}$ such that $|\bar{d}\bar{b}| = |db|$.

Such a neighborhood is called a *normal region.* One can always choose a normal region G so small that all shortest paths with endpoints in G are still contained in a (possibly larger) normal region.

Exercise 4.1.10. Check that Definition 4.1.9 is indeed a word-by-word reformulation of Definition 4.1.2 in new notation.

Exercise 4.1.11. Show that it is sufficient to require the triangle condition only for d being the midpoint of $[ab]$; i.e., such a weakened definition is still equivalent to the original one.

Exercise 4.1.12. Show that the following property is implied by but not equivalent to the triangle condition:

For any triangle $\triangle abc$, and midpoints d and e of its sides $[ab]$ and $[bc]$ the inequality $2|de| \le |ac|$ holds.

Hint: Consider a normed vector space.

Example 4.1.13. The direct metric product (see 3.6.1) of nonpositively (nonnegatively) curved spaces is a nonpositively (resp. nonnegatively) curved space.

Exercise 4.1.14. Prove this.

Hint: Let $\triangle_X$ be a triangle in $X = X_1 \times X_2$, $\triangle_1$ and $\triangle_2$ its projections to X_1 and X_2. Denote by $\triangle a_1b_1c_1$ and $\triangle a_2b_2c_2$ comparison triangles for $\triangle_X$ and $\triangle_Y$ respectively. Choose these triangles in the (x, y)- and (z, t)-planes

in Euclidean space $\mathbb{R}^4$ with Cartesian coordinates (x, y, z, t). Now observe that the triangle in $\mathbb{R}^2_1 \times \mathbb{R}^2_2$ with vertices (a_1, a_2), (b_1, b_2), and (c_1, c_2) is a comparison triangle for $\triangle_X$. Now a short straightforward computation finishes the proof.

Roughly speaking, all sufficiently small triangles in a space of nonpositive (resp. nonnegative) curvature are not thicker (resp. not thinner) than corresponding Euclidean triangles; in a nonnegatively curved space, a triangle may be thought of as "fat" (or "swollen") as in Figure 4.2, while a triangle in nonpositive curvature is skinny (with sides "sucked inside").

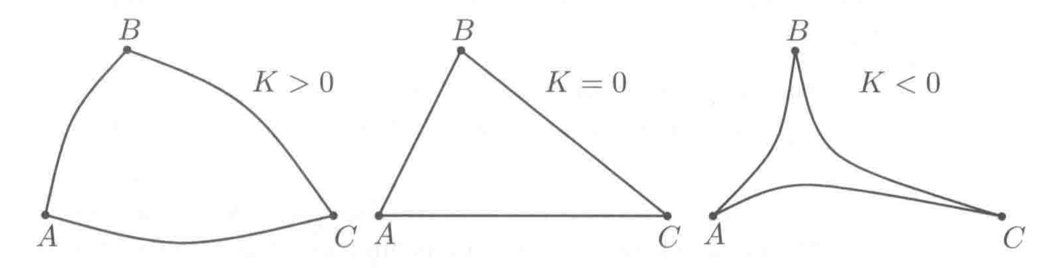


Figure 4.2: Comparison of triangles.

Though the definitions of nonpositively and nonnegatively curved spaces look similarly, we will see later that their properties are very different.

Please note that although one can speak about "a space *having* nonpositive (resp. nonnegative) curvature", we do not define a notion of curvature alone, and neither do we assign to it a numerical value.

4.1.5. Angle comparison for triangles. Looking at the figures of "fat" and "skinny" triangles, one naturally suspects that "fat" triangles should have large angles, while the angles of "skinny" triangles are expected to be small. Indeed, this observation leads to a definition of nonpositively curved spaces via comparisons of angles: X is a space of nonpositive curvature if the angles of every sufficiently small triangle exist and they are not greater than the corresponding angles of a comparison triangle in the Euclidean space.

In case of a nonnegatively curved space we have to change the words "not greater" to "not less" and *assume,* in addition, that the sum of adjacent angles is equal to π. [1] More formally, the new definition reads as follows.

Definition 4.1.15 ("*Angle condition*"). A length space X is a space of nonpositive curvature if every point of X has a neighborhood such that, for

[1] We do not know if the last condition is really necessary or not.

every triangle $\triangle abc$ contained in this neighborhood, the angles $\measuredangle bac$, $\measuredangle cba$ and $\measuredangle abc$ are well defined and satisfy the inequalities

$$\measuredangle bac \le \widetilde{\measuredangle} bac \quad \measuredangle abc \le \widetilde{\measuredangle} abc \quad \measuredangle bca \le \widetilde{\measuredangle} bca$$

(recall that $\widetilde{\measuredangle} abc$ denotes the comparison angle, i.e., $\widetilde{\measuredangle} abc = \measuredangle \bar{a}\bar{b}\bar{c}$ where $\triangle \bar{a}\bar{b}\bar{c}$ is a comparison triangle, cf. Definition 3.6.25).

A length space X is a space of nonnegative curvature if every point of X has a neighborhood such that, for every triangle $\triangle abc$ contained in this neighborhood, the angles $\measuredangle bac$, $\measuredangle cba$ and $\measuredangle abc$ are correctly defined and satisfy the inequalities

$$\measuredangle bac \ge \widetilde{\measuredangle} bac \quad \measuredangle abc \ge \widetilde{\measuredangle} abc \quad \measuredangle bca \ge \widetilde{\measuredangle} bca,$$

and, in addition, the following holds: for any two shortest path $[pq]$ and $[rs]$ where r is a inner point of $[pq]$, one has $\measuredangle prs + \measuredangle srq = \pi$.

While the triangle condition 4.1.9 is just a reformulation of Definition 4.1.2, proving their equivalence to the angle condition 4.1.15 requires some work. We prove the equivalence later in Section 4.3.2.

4.2. Examples

With the three definitions in mind, we are ready to give more examples of Alexandrov spaces. Apparently the examples that historically motivated the notion of curvature are convex and saddle surfaces. Unfortunately, we need to develop some machinery to prove that these examples satisfy our definition; however we suggest the reader keep them in mind.

Here we consider examples of Alexandrov spaces whose justification does not involve any techniques, namely a few trivial examples and polyhedral spaces in low dimensions (1 and 2).

Example 4.2.1. Euclidean spaces are obviously Alexandrov spaces (both nonpositively and nonnegatively curved at the same time).

Example 4.2.2. A convex set in an Alexandrov space is obviously an Alexandrov space (with the same sign of curvature).

Example 4.2.3. An open set in an Alexandrov space, regarded with the induced length metric, is an Alexandrov space (with the same sign of curvature). This is so because our definitions are local, and the induced length metric of an open set locally coincides with the metric restricted from the ambient length space.

Example 4.2.4. A "fan" made of several segments glued together at one end is a space of nonpositive curvature. The proof could be the same as for Example 4.1.3. But now, using the angle definition, we are able to prove

this in a few words: every triangle in our space either has three zero angles (in which case the angle condition obviously holds) or is degenerate (that is, contained in one of its sides). In the latter case its comparison triangle is degenerate as well.

Since every point in a locally-finite graph has a neighborhood that is either a segment or a bunch of segments attached to one point, all locally-finite connected graphs are nonpositively curved spaces.

Warning: Unlike graphs, not all two-dimensional locally-finite polyhedra are Alexandrov spaces.

Example 4.2.5. The union of the xy-plane and the z-axis in $\mathbb{R}^3$ (with the induced intrinsic metric) is another example of nonpositively curved space. This example can be obtained (and this is a better "intrinsic" description) by attaching together a plane and a line by gluing them at one point, that is, forming a metric bouquet (see Definition 4.2.7 below).

Proof. We will show that every triangle in our space satisfies the angle condition. Note that every shortest path with endpoints in the plane is entirely contained in the plane. And a shortest path starting at a point in the line can leave the line only through the origin O. Now one can see that the angle condition for a triangle is trivial if all its vertices belong to the plane or to the line, or if two vertices are in the line and one is in the plane. In the only nontrivial case when vertices a, b reside in the plane and c in the line, the triangle $\triangle abc$ consists of the flat triangle $\triangle abO$ and the "tail" Oc. It is then an elementary exercise to check that the angles of this triangle are less than the corresponding angles of a comparison triangle. □

Example 4.2.6. Consider several copies of the plane $\mathbb{R}^2$ and identify their origins. This gives a nonpositively curved space. For two copies of $\mathbb{R}^2$, this space topologically looks like the cone $x^2 + y^2 = z^2$.

Warning: This cone together with its intrinsic metric induced from $\mathbb{R}^3$ is *not* an Alexandrov space!

The proof is essentially the same as in the previous example.

Generalizing the above examples, let us give the following

Definition 4.2.7. Let $\{X_i\}$ be a collection of length spaces and a point x_i is chosen in every space X_i. The *metric bouquet* of spaces X_i (with marked points x_i) is a length space obtained from the disjoint union $\bigcup X_i$ by gluing all points x_i together.

Exercise 4.2.8. Prove that gluing length spaces into a bouquet does not change their metrics. In other words, the spaces X_i are projected isometrically onto their images in the bouquet.

The above three examples are partial cases of the following general statement.

Proposition 4.2.9. *A metric bouquet of nonpositively curved spaces is a nonpositively curved space.*

Exercise 4.2.10. Prove the proposition.

Example 4.2.11. "Notebook" example 2.2.7 from Section 2.2 (several half-planes glued together along their edges) is a nonpositively curved space.

Proof. Consider a triangle $\triangle abc$ in the "notebook" space X. The only interesting case is when the vertices a, b, c of the triangle lie in different half-planes, say, in half-planes A, B and C respectively. Let L be the common edge of the half-planes. The union of the half-planes $A \cup B$ is a Euclidean plane, and the shortest path $[ab]$ is a segment in this plane (this shortest path cannot visit the interior of C: such a visit makes a path longer because in this case $[ab]$ intersects L at least twice and an interval of the path in C is longer that the corresponding segment of L). Shortest paths $[bc]$ and $[ac]$ have similar properties. Now consider the isometry f of B to C fixing L. (Imagine that we turn B around L to C till they coincide.) Denote by d the intersection of $[bc]$ and L. Then the quadrangle $(a, b, d, f(c))$ has the same angles at points a, b, $f(c)$ as the triangle $\triangle abc$. It is easy to check (now it is elementary Euclidean geometry) that angles of the quadrangle are less than the angles of a comparison triangle for $\triangle abc$. □

Remark 4.2.12. In Chapter 9 we will prove a general Reshetnyak's theorem (Theorem 9.1.21), which covers the above nonpositively curved examples as well as many other spaces obtained by gluing.

It turns out that upper bounds for curvature are less restrictive and therefore there are fewer different types of examples of nonnegatively curved spaces. In particular, all two-dimensional spaces of nonnegative curvature are topological manifolds, possibly with boundary. We will see later that all convex surfaces are nonnegatively curved spaces. Here we prove this for polyhedral surfaces. More generally, we can carry out a complete analysis of which two-dimensional polyhedral spaces are Alexandrov spaces. It is advisable to refresh the basics of polyhedral metrics given in 3.2. We need the following important lemma, whose proof is already contained in the proof for Example 4.1.4 above.

Lemma 4.2.13. *The cone over a circle is a nonpositively curved space iff the length of the circle is at least 2π. The cone over a circle is a nonnegatively curved space iff the length of the circle is less than or equal to 2π.*

Using this lemma, we can characterize all two-dimensional polyhedral Alexandrov spaces. Recall that polyhedral length spaces were defined and considered in 3.2. Each point of a two-dimensional polyhedral space has a neighborhood isometric to a neighborhood of the vertex in a cone over a graph. This graph is called the *link* of the point.

Theorem 4.2.14. *1. A two-dimensional polyhedral space is a space of nonnegative curvature iff it is a topological manifold (possibly with boundary), and the sum of angles around every vertex is not greater than 2π, and moreover is not greater than π if the vertex belongs to the boundary. Equivalently, this means that the link of every vertex is a circle of length at most 2π or a segment of length at most π.*

2. A two-dimensional polyhedral space is a space of nonpositive curvature iff the link of each vertex does not contain a subspace isometric to a circle of length less than 2π.

Proof. Since being nonnegatively (resp. nonpositively) curved is a local property, it suffices to consider only a small neighborhood of each point. If the neighborhood is small enough, then there is no difference between the polyhedron and the cone over the link. In other words, a polyhedral space X is nonnegatively (resp. nonpositively) curved if and only if, for every $x \in X$, the cone over the link of x is a nonnegatively (resp. nonpositively) curved space.

1. A cone over a circle of length no greater than π is nonnegatively curved by Lemma 4.2.13. A cone over a segment of length $\alpha \leq \pi$ is isometric to a planar sector of angular measure α; this sector is nonnegatively curved as a convex subset of the plane. Thus, if the link of every point is a circle of length at most 2π or a segment of length at most π, then the space is nonnegatively curved.

Now assume that the space is nonpositively curved, and let us rule out all cases except these two types of links.

(a) Suppose that the link of some point is not connected. Then removal of this point vertex makes X locally disconnected. Considering a triangle $\triangle abc$ where a belongs in one component and b, c in another one, one can see that X cannot have nonnegative curvature (compare with Example 4.2.6).

(b) Suppose that more than two faces are adjacent to some edge. Then a neighborhood of a point in that edge looks like the "notebook" in Example 4.2.11, which is not nonnegatively curved.

(c) Now we know that the link is connected, and that every point in the link has degree at most 2 (otherwise in X there are more than two faces adjoint to one edge). Therefore the link is either a segment or a circle. If it

is a circle, Lemma 4.2.13 implies that its length is not greater than 2π. If the link is a segment, its length is not greater than π by Exercise 4.1.7.

2. Now we pass to the case of nonpositive curvature. We have to prove the following:

The cone K over a graph Γ is a space of nonpositive curvature if and only if the length of each nontrivial loop in Γ is not less than 2π.

First assume that K is nonpositively curved, and let us show that Γ does not contain nontrivial loops of length less than 2π. Let γ be a shortest nontrivial loop in Γ. Then, for every two points $x, y \in \gamma$, one of the two intervals of γ between x and y is a shortest path in Γ. It follows that γ is a convex set in Γ; hence the sub-cone over γ is a convex set in K (by Lemma 3.6.15). Therefore this sub-cone is nonpositively curved as long as the cone is; hence $L(\gamma) \geq 2\pi$ by Lemma 4.2.13. Since γ is a shortest loop, all loops have length no less than 2π.

Now assume that all nontrivial loops in Γ are not shorter than 2π, and let us prove that K is nonpositively curved. Consider a triangle $\triangle abc$ in K. First suppose that the sides of the triangle do not pass through O. Consider the projection of $\triangle abc$ to Γ. This projection is a triangle $\triangle a'b'c'$ whose sides are shortest paths in Γ (recall the discussion of shortest path in cones in Section 3.6.2). Consider two cases.

(a) Suppose that $\triangle a'b'c'$ (as a subset of Γ) does not contain a simple loop (a simple loop is a set homeomorphic to the circle). Then it is easy to see that all three sides of $\triangle a'b'c'$ have a common point $d \in \Gamma$; more precisely, there are shortest paths $[a'd]$, $[b'd]$ and $[c'd]$ such that the triangle $\triangle a'b'c'$, as a subset of Γ, coincides with the "fan" $[a'd] \cup [b'd] \cup [c'd]$. This fan is a convex set in Γ (because every pair of its points belong to a shortest path contained in the fan). Hence the original triangle $\triangle abc$ is contained in the cone over the fan which is a convex set in K consisting of three sectors glued together along a ray. Then absolutely the same argument as for the "notebook" Example 4.2.11 finishes the proof.

(b) Now suppose that $\triangle a'b'c'$ does contain a simple loop. Then the perimeter $L = |a'b'| + |b'c'| + |a'c'|$ is no less than 2π. Consider the cone K_1 over the circle S of length L. Fix a length-preserving map g from S onto the triangle $\triangle abc$; namely, split S into three arcs of lengths $|a'b'|$, $|b'c'|$ and $|a'c'|$, and let g map these arcs onto the respective sides of the triangle. This map g induces the map $\bar{g}\colon K_1 \to K$ sending a point $(x,t) \in K_1$ to the point $(g(x), t) \in K$. The map $\bar{g}$ is an arcwise isometry (and hence is nonexpanding); and the triangle $\triangle abc$ is the image of some triangle $\triangle a''b''c''$ in K_1 (again, recall the structure of shortest paths in cones discussed in Section 3.6.2).

Since the length of S is not less than 2π, K_1 is a space of nonpositive curvature. Since the triangles $\triangle abc$ in K and $\triangle a''b''c''$ in K_1 have equal lengths of sides, they have a common comparison triangle $\triangle\bar{a}\bar{b}\bar{c}$ in R^2. Then one has $|ad| \le |a''d''| \le |\bar{a}\bar{d}|$ for every point $d \in [bc]$ and corresponding points $d'' \in [b''c'']$ and $\bar{d} \in [\bar{b}\bar{c}]$. Here the inequality $|ad| \le |a''d''|$ follows from the fact that $\bar{g}$ is a nonexpanding map, and the inequality $|a''d''| \le |\bar{a}\bar{d}|$ from the fact that K_1 is nonpositively curved. This finishes the proof in the case (b).

It remains to consider the case when at least one side of $\triangle abc$ passes through the origin O of the cone. Suppose that the side $[ab]$ passes through O. Note that in this case the distance in Γ between the projections of a and b is greater than or equal to π.

If one or both of the sides $[bc]$ and $[ac]$ also pass through O, the proof is similar to Examples 4.2.5 and 4.2.4 and is left as an exercise. We consider only "the most general case" when neither $[ac]$ nor $[bc]$ passes through O. Similarly to the above, the projections of $[ac]$ and $[bc]$ to Γ are shortest paths $[a'c']$ and $[b'c']$ in Γ, and the triangle $\triangle abc$ is contained in the sub-cone over the union $[a'c'] \cup [b'c']$. Then, similarly to the case (b) above, the distance condition for the triangle $\triangle abc$ is reduced to that for a triangle in the cone K_1 over a segment of length $|a'c'| + |b'c'|$. Now the remaining part of the theorem follows from the fact that a cone over a segment is nonpositively curved (Exercise 4.1.7). □

4.3. Angles in Alexandrov Spaces and Equivalence of Definitions

4.3.1. Monotonicity of angles. Let us formulate one more definition of Alexandrov spaces. Then we will prove that all the definitions are equivalent.

Let α and β be two shortest paths (parameterized by arc length) starting at the same point p. Such configuration will be also referred to as "a hinge" α, β. As in Section 3.6.5, introduce the notation $\theta(x, y) = \tilde{\angle}\alpha(x)p\beta(y)$; i.e., $\theta(x, y)$ is the angle at $\bar{p}$ in a comparison triangle for $\triangle\alpha(x)p\beta(y)$.

Definition 4.3.1 ("*Monotonicity condition.*"). X is a space of nonnegative (resp. nonpositive) curvature if it can be covered by neighborhoods such that, for two any shortest segments α and β contained in this neighborhood (and starting from the same point p), the corresponding function $\theta(x, y)$ is nonincreasing (resp. nondecreasing) in each variable x and y (with the other one remaining fixed).

We have an immediate corollary:

Proposition 4.3.2. *If X is an Alexandrov space in the sense of Definition 4.3.1, then the angle between any two shortest paths in X is well-defined.*

4.3.2. Equivalence of the definitions. We need an elementary fact in Euclidean geometry, the so-called Alexandrov's lemma (this lemma was first used in a similar context in A. D. Alexandrov's book [**Al**]). Before giving its formal statement, let us explain its meaning from an engineer's viewpoint. Such associations may be very helpful in understanding geometric theorems.

Consider a plane quadrangle made of four rods connected by joints. Denote two opposite (diagonal) joints by b and d and assume that all angles of the quadrangle, possibly except at d, are less than π. Regarding the two sides attached to the joint b as a hinge, open this hinge up to straighten the hinge formed by the sides attached to d (if this is possible). If the angle at d were less than π (that is, if the quadrangle was convex), then this procedure would cause d to move closer to b, and it moves away from b otherwise.

Perhaps, drawing a figure may replace the proof and convince most readers.

Formally, it could be stated in this way:

Lemma 4.3.3 (Alexandrov's lemma). *Let a, b, c, d be points in the plane such that a and c are in different half-planes with respect to the line bd. Consider a triangle $\triangle a'b'c'$ in $\mathbb{R}^2$ such that*

$$|ab| = |a'b'|, \quad |bc| = |b'c'|, \quad |ad| + |dc| = |a'c'|$$

and let d' be a point in the side $[a'c']$ such that $|ad| = |a'd'|$.

Then $\measuredangle adb + \measuredangle bdc < \pi$ if and only if $|b'd'| < |bd|$. In this case, one also has $\measuredangle b'a'd' < \measuredangle bad$ and $\measuredangle b'c'd' < \measuredangle bcd$; see Figure 4.3.2,on left.

And $\measuredangle adb + \measuredangle bdc > \pi$ if and only if $|b'd'| > |bd|$, in which case one has $\measuredangle b'a'd' > \measuredangle bad$ and $\measuredangle b'c'd' > \measuredangle bcd$; see Figure 4.3.2,on right.

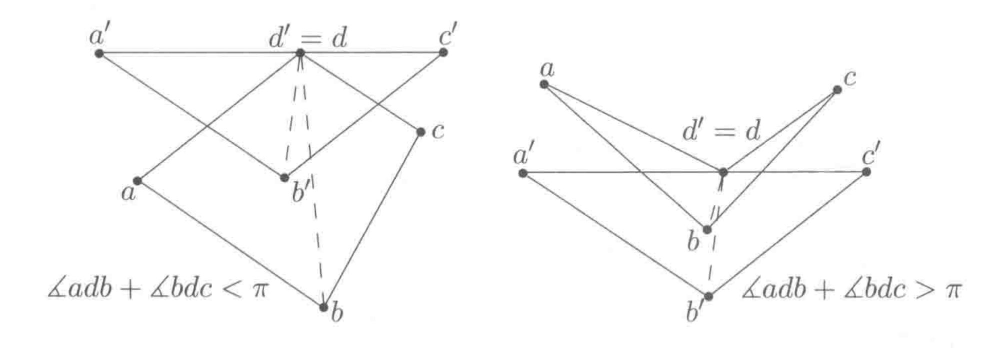


Figure 4.3: Distance bd decreases or increases depending on $\measuredangle adb + \measuredangle bdc$.

Proof. We use only the following fact: if two sides of a planar triangle are fixed, then the angle between them is a monotone (increasing) function of

the third side. In other words, if $|xy| = |x'y'|$ and $|yz| = |y'z'|$ for two planar triangles $\triangle xyz$ and $\triangle x'y'z'$, then $\measuredangle xyz > \measuredangle x'y'z'$ if and only if $|xz| > |x'z'|$, and vice versa.

Take a point c_1 on the ray ad so that d is between a and c_1, and $|dc| = |dc_1|$. Suppose that $\measuredangle adb + \measuredangle bdc > \pi$; then $\measuredangle bdc_1 < \measuredangle bdc$. Hence $|bc_1| < |bc| = |b'c'|$ from the triangles $\triangle bdc$ and $\triangle bdc_1$. Now applying the same observation to the triangles $\triangle abc_1$ and $\triangle a'b'c'$ (for which $|ab| = |a'b'|$ and $|ac_1| = |a'c'|$), we obtain that $\measuredangle bac_1 < \measuredangle b'a'c'$. Therefore $|bd| < |b'd'|$ (from the triangles $\triangle bad$ and $\triangle b'ad'$).

The case $\measuredangle adb + \measuredangle bdc < \pi$ is similar, up to reversing the inequalities. □

Remark 4.3.4. The lemma is true (and the same proof works) if the triangles are placed in a sphere or a hyperbolic plane (the latter is defined in Chapter 5) instead of the Euclidean plane. We will use this remark later when we define more general curvature bounds.

Theorem 4.3.5. *All the definitions* (distance 4.1.2, triangle 4.1.9, angle 4.1.15, and monotonicity 4.3.1) *are equivalent.*

Remark 4.3.6. Our definitions are local. Their equivalence means that if one of the definitions holds in a region U, then others hold in some region V, possibly smaller than U. Nevertheless we will refer to all such regions as normal regions.

Proof of Theorem 4.3.5. The proofs for spaces of nonpositive and nonnegative curvature are similar up to reversing the inequalities. To avoid repetition, we prove the equivalence for nonpositively curved spaces and then indicate the necessary modifications for the case of nonnegative curvature.

(1) The distance and triangle conditions are obviously equivalent.

(2) Assume the triangle condition holds, and let us show that then the monotonicity condition holds as well. Consider a hinge of two shortest paths $\alpha = [p, a]$, $[p, b]$ and a point a_1 in α starting at p. Consider comparison triangles $\triangle \bar{p}\bar{a}\bar{b}$ and $\triangle \bar{p}\bar{a}_1\bar{b}$ for the triangles $\triangle pab$ and $\triangle pa_1b$. Let $\tilde{a}$ be a point in the side $\bar{p}\bar{a}_1$ such that $|\bar{p}\tilde{a}| = |pa|$. Then the triangle condition implies $|\tilde{a}\bar{b}| \geq |ab| = |\bar{a}\bar{b}|$. This means that $\measuredangle \bar{a}_1\bar{p}\bar{b} \geq \measuredangle \bar{a}\bar{p}\bar{b}$, which is the monotonicity of angles.

(3) The monotonicity condition 4.3.1 implies the angle condition 4.1.15. To prove this, consider a triangle $\triangle abc$. Let its sides $[ba]$, $[bc]$ be the shortest paths α, β, $\alpha(0) = \beta(0) = b$. Monotonicity of angles implies that

$$\measuredangle abc \equiv \measuredangle(\alpha, \beta) = \lim_{t \to 0} \theta(t, t) \leq \theta(|ab|, |bc|)$$

where θ is as in 4.3.1. Since $\theta(|ab|, |bc|) = \measuredangle \bar{a}\bar{b}\bar{c}$, the angle condition follows.

(4) It remains to prove that the angle condition 4.1.15 implies the triangle condition 4.1.9. Consider a triangle $\triangle abc$ and a point d in its side $[ac]$. Note that $\measuredangle bda + \measuredangle bdc \geq \measuredangle adc = \pi$ by the triangle inequality for angles. Place comparison triangles $\triangle\bar{a}\bar{b}\bar{d}$ and $\triangle\bar{c}\bar{b}\bar{d}$ in different half-planes with respect to the line $\bar{b}\bar{d}$ in $\mathbb{R}^2$. Then by the angle condition we have $\measuredangle\bar{a}\bar{d}\bar{b} + \measuredangle\bar{c}\bar{d}\bar{b} \geq \pi$. Now by Alexandrov's lemma 4.3.3 it follows that $|bd| = |\bar{b}\bar{d}| \leq |\bar{b}_1\bar{d}_1|$, where $\triangle\bar{a}_1\bar{b}_1\bar{c}_1$ is a comparison triangle for $\triangle abc$ and $\bar{d}_1$ is the point in $[\bar{a}_1\bar{c}_1]$ such that $|\bar{a}_1\bar{d}_1| = |ad|$. This is the triangle condition for $\triangle abc$ and d, and the equivalence of the definitions (of nonpositive curvature) follows.

To prove the theorem for nonnegatively curved spaces, just reverse all inequalities. Besides inequalities originating from the definitions of bounded curvature, we used the fact that the sum of adjacent angles is not less than π. This is true in any length space but the opposite inequality does not hold automatically. However the requirement that the sum of adjacent angles equals π was included in the angle condition for nonnegative curvature (recall Definition 4.1.15). To complete the proof of the theorem, we have to show that this property follows from the monotonicity condition. We formulate this as a separate lemma (it has numerous applications).

Lemma 4.3.7. *If a space X has nonnegative curvature in the sense of monotonicity definition 4.3.1, then, for any shortest path, the sum of adjacent angles is equal to π. In other words, if q_0 is an inner point of a shortest path p_0r_0 and q_0s_0 is a shortest path, then $\measuredangle p_0q_0s_0 + \measuredangle s_0q_0r_0 = \pi$.*

Proof. We have $\measuredangle p_0q_0s_0 + \measuredangle s_0q_0r_0 \geq \measuredangle p_0q_0r_0 \geq \pi$ by the triangle inequality for angles. To prove the opposite inequality, consider points p, s, r in the shortest paths $[p_0q_0]$, $[s_0q_0]$, $[r_0q_0]$, respectively. Place comparison triangles $\triangle\bar{p}\bar{q}\bar{s}$ and $\triangle\bar{s}\bar{q}\bar{r}$ on different sides of the line $\bar{q}\bar{s}$ in $\mathbb{R}^2$. And let $\triangle\bar{p}_1\bar{s}_1\bar{r}_1$ be a comparison triangle for the triangle $\triangle psr$. The monotonicity condition 4.3.1 implies that $\measuredangle\bar{s}\bar{p}\bar{q} \geq \measuredangle\bar{s}_1\bar{p}_1\bar{r}_1$. Then by Alexandrov's Lemma 4.3.3 the quadrangle $\bar{p}\bar{q}\bar{r}\bar{s}$ is convex, i.e., $\measuredangle\bar{p}\bar{q}\bar{s} + \measuredangle\bar{r}\bar{q}\bar{s} \leq \pi$. Passing to a limit as points p, r, s approach q we obtain the desired inequality $\measuredangle pqs + \measuredangle sqr \leq \measuredangle pqr \leq \pi$. □

With this lemma, the proof of Theorem 4.3.5 is finished. □

Exercise 4.3.8. Prove that our conditions of nonpositivity (resp. nonnegativity) of curvature are equivalent to the following: let $\triangle pqr$ be a (sufficiently small) triangle in X and $\triangle\bar{p}\bar{q}\bar{r}$ be its comparison triangle. Then for points p_1, r_1 in the sides $[pq]$, $[rq]$ and points $\bar{p}_1, \bar{r}_1$ in the sides $[\bar{p}\bar{q}]$, $[\bar{r}\bar{q}]$

such that $|qp_1| = |\bar{q}\bar{p}_1|$, $|qr_1| = |\bar{q}\bar{r}_1|$ the inequality $|p_1r_1| \le |\bar{p}_1\bar{r}_1|$ (resp. $|p_1r_1| \ge |\bar{p}_1\bar{r}_1|$) holds.

4.3.3. Semi-continuity of angles. In the plane, angles of a triangle depend continuously on its vertices. This is no longer true in general length spaces. However, angles in Alexandrov spaces enjoy certain semi-continuity properties.

Suppose that the sequences of shortest segments $[a_ib_i]$ and $[a_ic_i]$ converge uniformly to shortest paths $[ab]$ and $[ac]$, respectively. (All shortest segments are considered as paths parameterized with constant speed.) Note that the uniform convergence of paths is much stronger than its implication that $a_i \to a$, $b_i \to b$ and $c_i \to c$. Nevertheless, in general it is impossible to claim anything about relations between the angle $\measuredangle bac$ and the limit of the sequence $\measuredangle b_ia_ic_i$ even if this limit exists.

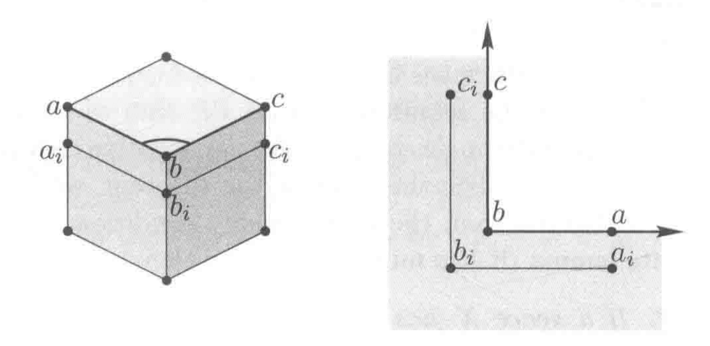


Figure 4.4: The limit angle can both collapse or explode.

Example 4.3.9. Let $[ab]$ and $[bc]$ be edges of a cube (we consider the surface of the cube but not the interior). Consider segments $[a_ib_i]$ and $[b_ic_i]$ parallel to $[ab]$ and $[bc]$ resp., at distances $1/i$ and placed in different faces of the cube; see Figure 4.4.

The hinges ($[a_ib_i]$, $[b_ic_i]$) converge to the hinge ($[ab]$, $[bc]$) as $i \to \infty$. However $\measuredangle abc = \pi/2$ while $\measuredangle a_ib_ic_i = \pi$ for all i.

Example 4.3.10. Consider the coordinate plane $\mathbb{R}^2$ without the coordinate quadrant $\{x > 0,\ y > 0\}$. Let $a = (1,0)$, $a_i = (1,-1/i)$, $b = (0,0)$, $b_i = (-1/i,-1/i)$, $c = (0,1)$, $c_i = (-1/i,1)$. Then the hinges ($[a_ib_i]$, $[b_ic_i]$) converge to the hinge ($[ab]$, $[bc]$), but $\measuredangle abc = \pi$ while $\lim_{i\to\infty} \measuredangle a_ib_ic_i = 1/2\pi$.

Note that the space in Example 4.3.9 has nonnegative curvature while the one in Example 4.3.10 has nonpositive curvature. This illustrates the following property of semi-continuity of angles

Theorem 4.3.11. *Let X be a space of nonpositive (resp. nonnegative) curvature. Suppose that the sequences of shortest paths $\{[a_ib_i]\}_{i=1}^{\infty}$ and $\{[a_ic_i]\}_{i=1}^{\infty}$ converge uniformly to shortest paths $[ab]$ and $[ac]$, respectively. Then $\angle bac \geq \limsup_{i\to\infty} \angle b_ia_ic_i$ (resp. $\angle bac \leq \liminf_{i\to\infty} \angle b_ia_ic_i$).*

Proof. Let α and α_i denote the angles $\angle bac$ and $\angle b_ia_ic_i$ respectively. For a small positive x, let $b' \in [a,b]$, $c' \in [a,c]$ and $b_i' \in [a_i,b_i]$, $c_i' \in [a_i,c_i]$ be points at the distance x from a and a_i respectively. Denote by $\theta(x)$ and $\theta_i(x)$ the comparison angles $\tilde{\angle} b'ac'$ and $\tilde{\angle} b_i'a_ic_i'$, respectively. Notice that $|a_ib_i'| = |ab|$, $|a_ic_i'| = |ac|$ and $|b_i'c_i'| \to |bc|$ as $i \to \infty$ (for fixed x), and hence $\lim_{i\to\infty} \theta_i(x) = \theta(x)$.

By the definition of the angle, $\alpha = \lim_{x\to 0} \theta(x)$ and $\alpha_i = \lim_{x\to 0} \theta_i(x)$. If X is nonpositively curved, then θ and θ_i are nondecreasing functions by the monotonicity condition. Therefore $\theta_i(x) \geq \alpha_i$ for all x, whence

$$\theta(x) = \lim_{i\to\infty} \theta_i(x) \geq \limsup_{i\to\infty} \alpha_i.$$

It follows that $\alpha = \lim_{x\to 0} \theta(x) \geq \limsup_{i\to\infty} \alpha_i$.

If X is nonnegatively curved, the proof is similar: θ_i is a nonincreasing function, hence $\theta_i(x) \leq \alpha_i$, then $\theta(x) \leq \liminf_{i\to\infty} \alpha_i$ for all x, and therefore $\alpha = \lim_{x\to 0} \theta(x) \leq \liminf_{i\to\infty} \alpha_i$. □

4.4. Analysis of Distance Functions

This section is optional, and nothing in the book relies on it. We already mentioned that distance functions in Alexandrov spaces curvature are "more convex" or "more concave" than those in the Euclidean plane (for nonpositive and nonnegative curvature, respectively). Here we put a formal wrapping around this claim. This approach to the definition of Alexandrov spaces is analytic rather than geometric, but it proves very useful in some cases.

As in Definition 4.1.2, consider a length space (X, d), a point $p \in X$, and a unit-speed shortest path $\gamma\colon [a,b] \to X$, and introduce the function $g(t) = d(\gamma(t), p)$ and the corresponding comparison function g_0 for the Euclidean comparison segment. Note that the function g_0 is uniquely determined by the function g; it is an easy trigonometric exercise to find a precise expression for g_0 in terms of $A = d(\gamma(a), p) = g(a)$ and $B = d(\gamma(b), p) = g(b)$.

Hence, we defined spaces of bounded curvature by restricting the class of functions that may arise as restrictions of distance functions to shortest paths. Let us make this description more explicit.

Notice that even without any curvature restrictions not every continuous function can arise as a 1-dimensional distance function. First of all, g obviously must be nonnegative; furthermore, it must be a nonexpanding

function. (Recall that a function g is called *nonexpanding* if $|g(t) - g(s)| \leq |t - s|$ for all s, t. For a smooth g, this is equivalent to the statement that $|g'(t)| \leq 1$ for all t.)

Exercise 4.4.1. Prove that g is nonexpanding.

Hint: This is a trivial consequence of the triangle inequality.

We need a complete list of all possible functions g_0 in one variable that can be obtained by restriction of the Euclidean distance function to a segment (in other words, the list of all functions we compare with). If $\bar{p} \in \mathbb{R}^2$ and γ_0 is a straight line in $\mathbb{R}^2$ parameterized with unit speed, then $g_0(t) = |\bar{p} - \gamma_0(t)| = \sqrt{(t+c)^2 + h^2}$ where c is the parameter such that $\gamma_0(-c)$ is the orthogonal projection of $\bar{p}$ to γ_0, and h is the distance from $\bar{p}$ to this projection. Thus the set of 1-dimensional Euclidean distance functions is the set of functions of the form $t \mapsto \sqrt{(t+c)^2 + h^2}$, where c and h are arbitrary real constants.

To develop a convenient language for expressing comparison inequalities, let us give a general definition.

Definition 4.4.2. Let F be a class of continuous functions. A continuous function $g\colon \mathbb{R} \to \mathbb{R}$ is called F-convex (resp. F-concave) if for every $f \in F$ such that $g(x) = f(x)$ and $g(y) = f(y)$, one has $f(z) \geq g(z)$ (resp. $f(z) \leq g(z)$) for all $z \in [x, y]$.

Example 4.4.3. For $F = \{f(t) = at + b, \quad a, b \in \mathbb{R}\}$ consisting of all linear functions, we get the usual notions of convex and concave functions.

Example 4.4.4. For $F_k\lambda = \{f(t) = \lambda t^2 + at + b, a, b \in \mathbb{R}\}$, we get the notion of λ-convexity. This notion means that one can touch the graph of g from below by a translation of the parabola $y = \lambda x^2$. In other words, while smooth convex functions are characterized by the inequality $g'' \geq 0$, for λ-convex functions this inequality turns into $g'' \geq \lambda$.

We are certainly concerned with the class of functions

$$E = \{f(t) = \sqrt{(t+c)^2 + h^2}, \;\; c, h \in \mathbb{R}\}.$$

Indeed, the additional restriction imposed by the distance condition 4.1.2 is that all 1-dimensional distance functions g are E-convex (resp. concave).

Remark 4.4.5. To be more precise, our requirement is assumed to hold only locally. Thus, for a 1-dimensional distance function g, we should require that the restriction of g to an interval is E-convex (or E-concave) if the maximum value of this restriction is sufficiently small. For the sake of simplicity we ignore this circumstance and consider E-convex functions $\mathbb{R} \to \mathbb{R}$.

The next exercise is an analog of the fact that a convex function possesses a "supporting linear function" at every point.

Exercise 4.4.6. Prove that a nonnegative nonexpanding function $g\colon \mathbb{R} \to \mathbb{R}$ is E-convex (resp. E-concave) if and only if for every $x \in \mathbb{R}$ there exists a function $f \in E$ such that $f(x) = g(x)$ and $f \le g$ (resp. $f \ge g$) everywhere.

Notice that in particular all nonnegative nonexpanding E-convex functions are convex in the usual sense.

The class E is the set of nonnegative solutions of the differential equation $g''(t)g(t) = 1 - (g'(t))^2$. Analogously to the case of convex functions, a smooth nonnegative nonexpanding function g is E-convex (resp. E-concave) if and only if

$$g''(t) \ge \frac{1-(g'(t))^2}{g(t)} \qquad (\text{resp. } g''(t) \le \frac{1-(g'(t))^2}{g(t)}\ ).$$

Exercise 4.4.7. Prove this.

Convex functions, possibly nonsmooth, enjoy nice properties: they have right and left derivatives everywhere; they are differentiable except at no more than countably many points, and their derivatives have positive (resp. negative) jumps at points of nonsmoothness. The same is true for E-convex and E-concave functions:

Exercise 4.4.8. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a nonnegative nonexpanding E-convex (resp. E-concave) function. Prove that

1. g is continuous.

2. g has right and left derivatives everywhere, and the left derivative is not greater (resp. not less) than the right one.

3. The set of points where g is not differentiable (i.e., where the right and left derivatives are not equal) is finite or countable.

4. The derivative of g is continuous on the set where it is defined.

Hint: $g(t)^2 - t^2$ is a convex (resp. concave) function.

4.5. The First Variation Formula

The term "first variation formula" comes from differential geometry and means a rule for differentiating the length of a variable curve. In this section we prove such a formula for shortest paths in Alexandrov spaces; furthermore, we restrict ourselves to the case when one endpoint of a shortest path is fixed and the other one is moving along a geodesic. (For more general cases, see exercises in the end of the section.) Since the length of a shortest path equals the distance between its endpoints, this also gives us a rule for differentiating distance functions in Alexandrov spaces.

Let us first consider the Euclidean case. Let $\gamma\colon [0,a] \to \mathbb{R}^2$ be a smooth unit-speed curve in $\mathbb{R}^2$ and $p \in \mathbb{R}^2$ a point not belonging to (the image of) γ. Consider the distance function $l(t) = |p\gamma(t)|$. Then

$$\frac{dl}{dt} = -\cos\measuredangle(p-\gamma(t),\ \gamma'(t)),$$

where γ' is, of course, the velocity of γ.

Exercise 4.5.1. Prove this formula.

We are going to show that similar formulas are valid for spaces of nonpositive and nonnegative curvature.

Later on we use the following notation. Let X be a length space, $\gamma\colon [0,T] \to X$ a unit-speed shortest path, $a = \gamma(0)$, $d = \gamma(T)$, and $p \in X \setminus \{a\}$. For each $t \in [0,T]$, set $l(t) = |p\gamma(t)|$ and fix a shortest path σ_t connecting $\gamma(t)$ to p.

The following proposition is a general fact; it is valid for any length space without any curvature restrictions.

Proposition 4.5.2. *If there exist the angle $\alpha = \measuredangle pad$ between the shortest paths γ and $[ap] = \sigma_0$, then*

$$\limsup_{t\to+0} \frac{l(t)-l(0)}{t} \le -\cos\alpha. \tag{4.1}$$

Remark 4.5.3. The reader who likes the traditional notation for infinitesimal quantities will probably prefer the following form of the inequality (4.1):

$$l(t) \le l(0) - t\cos\alpha + o(t), \qquad t \to +0.$$

Remark 4.5.4. Since the left-hand side of (4.1) does not depend on σ_0, one can write

$$\limsup_{t\to 0} \frac{l(t)-l(0)}{t} \le \inf_{\sigma_0}(-\cos\alpha),$$

or, equivalently,

$$\limsup_{t\to 0} \frac{l(t)-l(0)}{t} \le -\cos\alpha_{\min}$$

where $\alpha_{\min}$ is the infimum of angles between γ and all possible shortest paths from a to p. (See also Corollary 4.5.7 below.)

We need the following elementary

Lemma 4.5.5. *Let $\triangle abc$ be a triangle in $\mathbb{R}^2$, $\alpha = \measuredangle bac$, $t = |ac|$. Then*

$$\left|\cos\alpha - \frac{|ab|-|bc|}{t}\right| \le \frac{t}{|ab|}.$$

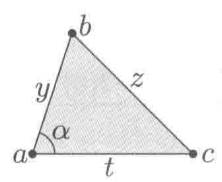


Figure 4.5: Lemma 4.5.5.

Proof. Denote $|ab| = y$ and $|bc| = z$ (see Figure 4.5). By the cosine rule we have

$$\cos\alpha = \frac{t^2 + y^2 - z^2}{2ty} = \frac{y-z}{t}\frac{y+z}{2y} + \frac{t}{2y}.$$

Then

$$\begin{aligned}\left|\cos\alpha - \frac{y-z}{t}\right| &= \left|\frac{y-z}{t}\frac{y+z}{2y} + \frac{t}{2y} - \frac{y-z}{t}\right| \\ &\le \left|\frac{y-z}{t}\right| \cdot \left|\frac{y+z}{2y} - 1\right| + \frac{t}{2y} \le 1 \cdot \frac{t}{2y} + \frac{t}{2y} \le \frac{t}{y}\end{aligned}$$

because $\left|\frac{y-z}{t}\right| \le 1$ and $\left|\frac{y+z}{2y} - 1\right| \le \frac{t}{2y}$ by the triangle inequality. □

Proof of Proposition 4.5.2. Consider two "variable" points: a point b in the shortest path $[ap] = \sigma_0$ and a point $c = \gamma(t)$. The triangle inequality implies that

$$|ab| - |bc| = |ap| - (|bp| + |bc|) \le l(0) - l(t).$$

Then apply Lemma 4.5.5 to the comparison triangle for $\triangle abc$. This yields

$$\cos \widetilde{\measuredangle} abc \le \frac{|ab| - |bc|}{t} + \frac{t}{|ab|} \le -\frac{l(t) - l(0)}{t} + \frac{t}{|ab|}.$$

One can let points b and c converge to a so that $t/|ab| \to 0$. Then the proposition follows by passing to the limit in the last inequality. □

The following theorem is the main result of this section. We keep on using the notation introduced before Proposition 4.5.2.

Theorem 4.5.6 (First variation theorem)**.** *Let X be a space of nonpositive or nonnegative curvature, let γ, σ_t and $l(t)$ be as above, and assume that a sequence σ_{t_i} converges to σ_0 for some sequence $\{t_i\}_{i=1}^{\infty}$, $t_i \to 0$ as $i \to \infty$. Then there exists a limit*

$$\lim_{t_i \to 0} \frac{l(t_i) - l(0)}{t_i} = -\cos\alpha, \tag{4.2}$$

where α is the angle at a between σ_0 and γ.

Proof of Theorem 4.5.6. By Proposition 4.5.2 we have

$$\limsup_{i\to\infty} \frac{l(t_i)-l(0)}{t_i} \le -\cos\alpha.$$

Hence it suffices to prove that

$$\liminf_{i\to\infty} \frac{l(t_i)-l(0)}{t_i} \ge -\cos\alpha.$$

Fix an $r>0$ so small that $|ap|>5r$ and the ball $B_{5r}(a)$ is a normal region for nonpositive or nonnegative curvature (this ball contains all triangles used in the constructions below). We may assume that $\gamma(t_i)\in B_r(a)$ for all i. For each i, set $c_i=\gamma(t_i)$ and let b_i be the point in the shortest path $[c_ip]=\sigma_{t_i}$ such that $|b_ic_i|=r$. We will prove that

$$\limsup_{i\to\infty} \widetilde{\measuredangle} ac_ib_i \le \pi-\alpha. \tag{4.3}$$

First let us show how the theorem follows from (4.3). Applying Lemma 4.5.5 to a comparison triangle for $\triangle ac_ib_i$ yields

$$l(0)=|pa|\le|pb_i|+|b_ia|\le|pb_i|+|b_ic_i|-t_i\cos\widetilde{\measuredangle} ac_ib_i+\frac{t_i^2}{|b_ic_i|}.$$

Since $|pb_i|+|b_ic_i|=l(t_i)$, it follows that

$$\frac{l(t_i)-l(0)}{t_i}\ge\cos\widetilde{\measuredangle} ac_ib_i-\frac{t_i}{|b_ic_i|}=\cos\widetilde{\measuredangle} ac_ib_i-\frac{t_i}{r}.$$

Then by (4.3),

$$\liminf_{i\to\infty}\frac{l(t_i)-l(0)}{t_i}\ge\liminf_{i\to\infty}(\cos\widetilde{\measuredangle} ac_ib_i)\ge\cos(\pi-\alpha)=-\cos\alpha,$$

and the theorem follows.

The proof for (4.3) is different for nonpositively and nonnegatively curved spaces.

1. Let X be a space of nonnegative curvature. Then

$$\widetilde{\measuredangle} ac_ib_i\le\measuredangle ac_ib_i=\pi-\measuredangle b_ic_id$$

by the angle condition. Then (4.3) follows because $\liminf_{i\to\infty}\measuredangle b_ic_id\ge\alpha$ by semi-continuity of angles (Theorem 4.3.11).

2. Let X be a space of nonpositive curvature. Denote by b the point in $[ap]=\sigma_0$ such that $|ab|=r$. Then $\measuredangle bab_i\le\widetilde{\measuredangle} bab_i$, and $\widetilde{\measuredangle} bab_i\to0$ as $i\to\infty$ because $|b_ib|\to0$ while $|ab|$ and $|ab_i|$ stay bounded away from zero. Hence $\measuredangle bab_i\to0$ as $i\to\infty$.

By the triangle inequality for angles it follows that $\measuredangle c_iab_i\to\alpha$ as $i\to\infty$. Then

$$\liminf_{i\to\infty}\widetilde{\measuredangle} c_iab_i\ge\liminf_{i\to\infty}\measuredangle c_iab_i=\alpha$$

by the angle condition. On the other hand, $\tilde{\angle} c_i a b_i + \tilde{\angle} a c_i b_i \to \pi$ as $i \to \infty$ because $\tilde{\angle} c_i a b_i + \tilde{\angle} a c_i b_i + \tilde{\angle} a b_i c_i = \pi$ and $\tilde{\angle} a b_i c_i \to 0$. Thus

$$\limsup_{i\to\infty} \tilde{\angle} a c_i b_i = \pi - \liminf_{i\to\infty} \tilde{\angle} c_i a b_i \le \pi - \alpha.$$

This proves (4.3) for nonpositive curvature. □

Theorem 4.5.6 obviously implies the following rule for differentiating the length: if $\{\sigma_t\}_{t\in[0,T]}$ is a continuous family of shortest paths connecting p to points $\gamma(t)$, then there exists the right derivative $dl/dt|_{t=0} = -\cos\alpha$. Moreover, one can differentiate the *distance* from p to $\gamma(t)$, even if shortest paths connecting p to $\gamma(t)$ are not unique. Namely, the following holds:

Corollary 4.5.7. *Let X be a nonpositively or nonnegatively curved complete locally compact space, $\gamma : [0,T]$ a geodesic parameterized by arc length, $p \in X$, $p \neq \gamma(0)$. Then the function $t \mapsto l(t) = |p\gamma(t)|$ has the right derivative and*

$$\lim_{t\to+0} \frac{l(t) - l(0)}{t} = -\cos\alpha_{\min}$$

where $\alpha_{\min}$ is the infimum (in fact, minimum) of angles between γ and shortest paths connecting $\gamma(0)$ to p.

Proof. By Proposition 4.5.2, we have

$$\limsup_{t\to+0} \frac{l(t) - l(0)}{t} \le -\cos\alpha_{\min}.$$

Choose a sequence $\{t_i\}$ converging to 0 such that

$$\frac{l(t_i) - l(0)}{t_i} \longrightarrow \liminf_{t\to+0} \frac{l(t) - l(0)}{t}$$

as $i \to \infty$, and fix shortest paths σ_{t_i} connecting p to $\gamma(t_i)$. By the Arzela–Ascoli Theorem 2.5.14, $\{\sigma_{t_i}\}$ contains a subsequence converging to some shortest path σ_0. We may assume that the sequence $\{\sigma_{t_i}\}$ itself converges to σ_0. Then by Theorem 4.5.6,

$$\lim_{i\to\infty} \frac{l(t_i) - l(0)}{t_i} = -\cos\alpha$$

where α is the angle between γ and σ_0. Thus

$$\liminf_{t\to+0} \frac{l(t) - l(0)}{t} = -\cos\alpha \ge -\cos\alpha_{\min}$$

and the formula for the right derivative follows. Note that the equality $\alpha = \alpha_{\min}$ also follows, so the minimum of angles is indeed attained (at σ_0). □

Remark 4.5.8. Theorem 4.5.6 and Corollary 4.5.7 imply a restriction on a shortest path σ_0 that can be obtained as a limit of shortest paths $\{\sigma_{t_i}\}$, $t_i \to 0$. Namely, if σ_0 is such a limit, then the angle $\alpha = \measuredangle(\gamma, \sigma_0)$ equals $\alpha_{\min}$ from Corollary 4.5.7. Indeed, the limit $(l(0) - l(t_i))/t_i$ equals $\cos\alpha$ by the theorem, and the same limit equals $\cos\alpha_{\min}$ by the corollary. Therefore $\alpha = \alpha_{\min}$.

Exercise 4.5.9. Prove that Theorem 4.5.6 is valid for an arbitrary curve γ (not necessarily a shortest path) having a direction at the initial point a.

Exercise 4.5.10. Generalize Theorem 4.5.6 to the case when both endpoints of shortest paths $\{\sigma_t\}$ are variable and move along two geodesics γ_1 and γ_2 with constant speeds v_1 and v_2. Namely prove that, if σ_t denotes a shortest path connecting $\gamma_1(t)$ and $\gamma_2(t)$ and a sequence $\{\sigma(t_i)\}$ converges to σ_0, then

$$\lim_{i\to\infty} \frac{l(t_i) - l(0)}{t_i} = -v_1 \cos\alpha_1 - v_2 \cos\alpha_2$$

where $l(t) = L(\sigma_t) = |\gamma_1(t)\gamma_2(t)|$ and $\alpha_j = \cos\measuredangle(\sigma_0, \gamma_j)$ for $j = 1, 2$.

Exercise 4.5.11. Generalize Corollary 4.5.7 to distance functions of sets. Namely prove the following. If $\gamma\colon [0,T] \to X$ is a shortest path, A is a closed set not containing $\gamma(0)$, and $l(t) = \operatorname{dist}(\gamma(t), A)$, then

$$\lim_{t\to+0} \frac{l(t) - l(0)}{t} = -\cos\alpha_{\min}$$

where $\alpha_{\min}$ is the minimum of angles between γ and shortest paths of length $l(0)$ connecting $\gamma(0)$ to A.

Remark 4.5.12. Theorem 4.5.6 is valid for general Alexandrov spaces of curvature bounded below or above (defined in Section 4.6), and the same proof works with minimal modifications.

4.6. Nonzero Curvature Bounds and Globalization

4.6.1. Nonzero curvature bounds. So far, we considered only spaces of nonpositive or nonnegative curvature. This was a simplification intended to help the reader understand the subject. Now we are going to define Alexandrov spaces of curvature not greater than k and no less than k, for an arbitrary $k \in \mathbb{R}$. These generalizations are not just a pursuit of generality. The reason is that spaces of curvature, say, ≥ 1 or ≤ -1, enjoy new important properties for which comparison with zero curvature is not sufficient.

In fact, it is sufficient to consider only two cases, $k = 1$ and $k = -1$, in addition to the case $k = 0$, because all other cases may be reduced to these

three by re-scaling. We still assume that all metrics in question are strictly intrinsic (more general definitions can be found in Chapters 9 and 10).

Historically, the notion of curvature comes from differential geometry in the form of Gaussian curvature for two-dimensional surfaces or Riemannian manifolds, and sectional curvatures in higher dimensions. (At this point, it does not matter what these terms mean; they are just classical objects of differential geometry. We will consider Riemannian manifolds and their curvatures in Chapters 5 and 6. For now, it suffices to mention that Gaussian and sectional curvatures are real-valued functions defined by means of certain differential expressions.) A Riemannian manifold is a nonpositively (resp. nonnegatively) curved Alexandrov space if and only if its sectional curvatures are nonpositive (resp. nonnegative) everywhere. These two classes of Riemannian manifolds play an important role in Riemannian geometry, and this is one of the motivations for studying Alexandrov spaces.

Similarly, Alexandrov spaces of curvature $\geq k$ and $\leq k$ (that we are about to define) include all Riemannian manifolds whose sectional curvatures are $\geq k$ (resp. $\leq k$) everywhere.

To stress the connection between our previous classes (of nonnegative and nonpositive curvature) and new ones (not defined yet), we mention the following facts that will be proved later:

1. If X is a length space of nonnegative curvature (or, more generally, of curvature bounded below), then its spaces of directions are spaces of curvature ≥ 1.

2. If X is a length space of nonpositive curvature (or, more generally, of curvature bounded above), then its spaces of directions have curvature ≤ 1.

Definitions. The essence of the generalization is that we compare with other "model spaces" instead of the Euclidean plane. In fact, these model spaces are standard two-dimensional spaces of constant Gaussian curvature k. For $k > 0$, the model space is the Euclidean sphere of radius $1/\sqrt{k}$. For $k < 0$, the model space is the hyperbolic plane of curvature k. Hyperbolic planes are defined later in Chapter 5. For now, you may restrict yourself to the case $k \geq 0$; then revisit this section after getting familiar with hyperbolic planes.

To avoid considering the three cases $k < 0$, $k = 0$ and $k > 0$ separately, we introduce the notion of k-plane:

Definition 4.6.1. Let k be a real number. The *k-plane* is one of the following spaces, depending on the sign of k:

- $\mathbb{R}^2$, if $k = 0$;

- the Euclidean sphere of radius $1/\sqrt{k}$ (with its intrinsic metric), if $k > 0$;
- the hyperbolic plane of curvature k, that is, the standard Lobachevsky plane with the metric multiplied by $1/\sqrt{-k}$, if $k < 0$.

Note that the k-plane is bounded (i.e., has finite diameter) if $k > 0$, and not bounded if $k \le 0$. Denote the diameter of the k-plane by R_k, i.e.,

$$R_k = \begin{cases} \pi/\sqrt{k}, & k > 0, \\ \infty, & k \le 0. \end{cases}$$

We need the following elementary properties of k-planes: for any $a, b, c > 0$ such that $a + b + c < 2R_k$, there exists a triangle in the k-plane with sides a, b, c; moreover such a triangle is unique up to a rigid motion (that is, an isometry from the k-plane to itself). Therefore for every sufficiently small triangle in a length space, there is a unique (up to a rigid motion) comparison triangle in the k-plane. (The words "sufficiently small" here can be omitted if $k \le 0$.)

Now one can define spaces of curvature $\ge k$ and of curvature $\le k$ in the same words as we defined nonnegatively and nonpositively curved spaces, but with the k-plane instead of the plane.

For instance, here is the new "triangle definition":

Definition 4.6.2. Let k be a real number. A length space X (with a strictly intrinsic metric) is a space of curvature $\ge k$ (resp. $\le k$) if every point $x \in X$ has a neighborhood U such that for any triangle $\triangle abc$ containing in U and any point $d \in [ac]$ the inequality $|bd| \ge |\bar{b}\bar{d}|$ (resp., $|bd| \le |\bar{b}\bar{d}|$) holds, where $\triangle\bar{a}\bar{b}\bar{c}$ is a comparison triangle in the k-plane and $\bar{d} \in [\bar{a}\bar{c}]$ is the point such that $|\bar{a}\bar{d}| = |ad|$.

All the other definitions can be re-formulated in the similar way. Except a few inessential points, all statements and proofs related to the equivalence of the definitions and simplest properties remain the same as for $k = 0$. Later we will see that local properties of spaces of curvature $\le k$ and of curvature $\ge k$ (well, not all local properties but all that we are interested in) do not depend on k. But there are important global properties (properties "in the large") that do depend on k. More precisely, spaces of curvature $\ge k$ where $k > 0$ and spaces of curvature $\le k$ where $k < 0$ definitely have additional interesting nonlocal properties.

Exercise 4.6.3. Formulate distance, angle and monotonicity definitions of spaces of curvature $\ge k$ and $\le k$ for any k. Prove equivalence of these definitions.

Exercise 4.6.4. If $k_1 > k_2$, then every space of curvature $\geq k_1$ is a space of curvature $\geq k_2$, and every space of curvature $\leq k_2$ is a space of curvature $\leq k_1$. Prove this.

The following definition introduces Alexandrov spaces with "variable" curvature bounds.

Definition 4.6.5. A length space X is a space of curvature bounded above (resp. below) if every point $x \in X$ has a neighborhood which is a space of curvature $\leq k$ (resp. $\geq k$) for some $k \in \mathbb{R}$. (A neighborhood is regarded with its induced length metric; k may depend on x.)

4.6.2. Globalization. There is an important class of Alexandrov spaces for which triangle comparison works "in the large", i.e., for all triangles no matter how big they are.

Definition 4.6.6. We say that a length space X is a space of curvature $\geq k$ or $\leq k$ *globally*, or *in the large* if the triangle condition from Definition 4.6.2 is satisfied for all triangles $\triangle abc$ in X for which a comparison triangle in the k-plane is well-defined, i.e., exists and is unique up to a rigid motion.

In other words, having curvature $\leq k$ or $\geq k$ in the large means that the whole space is a normal region.

It may be unclear how to apply this definition to nonconnected length spaces (in which some distances are infinite). The answer is the same as with local definitions: a length space X has curvature $\geq k$ or $\leq k$ in the large if and only if every connected component of X is. Below we implicitly assume that the spaces under consideration are connected.

The condition that a comparison triangle is well-defined requires some clarification. The matter is that if $k > 0$, then in the k-plane there are no triangles with perimeter greater than $2R_k = 2\pi/\sqrt{k}$, so no comparison triangle exists if the perimeter of $\triangle abc$ is greater than $2\pi/\sqrt{k}$. If the perimeter equals $2\pi/\sqrt{k}$, then there are two cases:

(1) All sides are shorter than $\pi/\sqrt{k}$. In this case, a comparison triangle is unique: it is a great circle with three points marked as vertices.
(2) One of the sides equals $\pi/\sqrt{k}$, say $|ab| = \pi/\sqrt{k}$. In this case, there are many different comparison triangles. One can take two opposite points in the sphere for $\bar{a}$ and $\bar{b}$ and connect them by two arbitrary great half-circles, then place $\bar{c}$ in one of these half-circles and consider the other half-circle as the side $[\bar{a}\bar{b}]$.

Thus, the condition that a comparison triangle for $\triangle abc$ is well-defined is equivalent to the following system of inequalities: $\max\{|ab|, |ac|, |bc|\} < \pi/\sqrt{k}$ and $|ab| + |ac| + |bc| \leq 2\pi/\sqrt{k}$.

In fact, it is not necessary to consider triangles of perimeter $2\pi/\sqrt{k}$:

Exercise 4.6.7. Let $k > 0$, and X is a length space such that the distance condition from Definition 4.6.2 (either for curvature $\geq k$ or for curvature $\leq k$) holds for all triangles whose perimeters are less than $2\pi/\sqrt{k}$. Prove that X is a space of curvature respectively $\geq k$ or $\leq k$ in the large.

Hint: In the case of curvature $\leq k$, the distance condition for triangles of perimeter $2\pi/\sqrt{k}$ simply reduces to nothing. In the case of curvature $\geq k$, approximate a given triangle of perimeter $2\pi/\sqrt{k}$ by triangles with smaller perimeters and obtain the desired inequalities by passing to a limit.

Similarly to Definition 4.6.6, one can "globalize" other curvature comparison conditions, and the global definitions are equivalent just like local ones.

Exercise 4.6.8. Formulate the global versions of distance, angle and monotonicity definitions of spaces of curvature $\geq k$ and $\leq k$ for any k, and prove equivalence of these definitions.

Exercise 4.6.9. If $k_1 > k_2$, then every space of curvature $\leq k_2$ in the large is a space of curvature $\leq k_1$ in the large, and every space of curvature $\geq k_1$ in the large is a space of curvature $\geq k_2$ in the large. Prove this.

The reader probably has noticed that the second part of the above exercise is much harder than the first one. The reason is, of course, that in the case $k_1 > 0$ the definition of curvature bounded by k_2 involves more triangles than that of curvature bounded by k_1. Namely the latter does not tell us anything about triangles of perimeter greater than $2\pi/\sqrt{k_1}$ while the former requires considering triangles of perimeters up to the greater value $2\pi/\sqrt{k_2}$ (or infinity if $k_1 < 0$).

In fact, this difficulty does not exist: if a space has curvature $\geq k$ in the large, *then all triangles have perimeters no greater than* $2\pi/\sqrt{k}$. Well, this rule has some exceptions among one-dimensional spaces (intervals and circles). We will give a precise formulation and prove this fact in Chapter 10.

Globalization theorems. There are two theorems saying that, under some conditions, local curvature bounds imply global ones. These theorems are very important; actually they are central facts of the whole theory. Though we prove these theorems in later chapters, it is worthwhile to formulate them here.

1. Globalization theorem for nonpositive curvature (Theorem 9.2.9): *every complete simply connected space of curvature $\leq k$, where $k \leq 0$, is a space of curvature $\leq k$ in the large.*

2. Toponogov's globalization theorem (Theorem 10.3.1): *for any $k \in \mathbb{R}$, every complete space of curvature $\geq k$ is a space of curvature $\geq k$ in the large.*

You see that the second theorem is very general, while the first one includes two additional assumptions: namely, it does not work for positive k, and it requires that the space is simply connected. It is easy to see that the topological requirement is necessary; for example, the product $S^1 \times \mathbb{R}$ is nonpositively curved (because it is locally isometric to $\mathbb{R}^2$) but is not nonpositively curved in the large (consider a triangle with vertices in the circle $S^1 \times \{0\} \subset S^1 \times \mathbb{R}$).

We suggest you prove these two theorems as exercises in the relatively simple case when $k = 0$ and the space is homeomorphic to $\mathbb{R}^2$. (If you wish, assume in addition that the space is polyhedral; however this does not simplify the proof much.)

4.7. Curvature of Cones

Here we add one more example to our list of Alexandrov spaces. Namely, we carry out a complete analysis of when a cone over a length space has nonpositive or nonnegative curvature. The cone over a metric space was defined in Section 3.6.2. Recall that points of the cone over a space X are pairs (x, t), where $x \in X$ and $t \in [0, +\infty)$, and all the pairs $(x, 0)$ are identified (the resulting point is called the origin, or the vertex, of the cone). The metric on the cone is defined by an analog of the Euclidean cosine law; in particular, the cone over the unit sphere S^2 is Euclidean space.

The sphere is the standard space of constant curvature 1, and Euclidean space has zero curvature (in our context, it is better to say that it is both nonpositively and nonnegatively curved). It is natural to expect that, if a length space X is more curved than the sphere (i.e., has curvature ≥ 1), then the cone over X is more curved than Euclidean space (i.e., has curvature ≥ 0), and vice versa. In fact, more assumptions about X are required to make this statement correct. The issue is that the definitions of bounded curvature are local, but the local structure of the cone near its vertex depends on the *global* structure of the base space X.

For instance, recall cones over circles (Example 4.1.4). Outside the origin, every cone over a circle is locally isometric to $\mathbb{R}^2$, but the geometry near the origin depends on the total length of the circle. In particular, cones over long circles are nonpositively (but not nonnegatively) curved, and cones over short circles are nonnegatively (but not nonpositively) curved.

For general cones, the following theorem holds:

Theorem 4.7.1. *Let K be a cone over a (possibly not connected) length space X and O its origin. Then*

1. $K \setminus \{O\}$ is a space of curvature ≥ 0 (resp. ≤ 0) if and only if X is a space of curvature ≥ 1 (resp. ≤ 1).

2. K is a space of curvature ≤ 0 if and only if X is a (possibly not connected) space of curvature ≤ 1 in the large.

3. K is a space of curvature ≥ 0 if and only if either X consists of exactly two points, or X is a connected space of curvature ≥ 1 in the large and no triangle in X has perimeter greater than 2π.

Remark 4.7.2. As we already mentioned, the last condition on the perimeters is satisfied automatically for "almost all" spaces of curvature ≥ 1 (more precisely, for all spaces except a few one-dimensional counter-examples).

This allows one to simplify the last statement of the theorem. We will discuss this further in Section 10.2.1.

Remark 4.7.3. A local curvature bound for K immediately implies the corresponding global one. To prove this, observe that any triangle in K can be moved to a normal neighborhood of O by means of re-scaling.

Proof of the theorem. 1. Consider a triangle $\triangle abc$ in K whose sides do not pass through O. Its projection to X is a triangle $\triangle a'b'c'$ in X with side lengths less than π. The sides of $\triangle abc$ are contained in convex flat sectors, namely in the sub-cones over the sides of $\triangle a'b'c'$. (See Figure 4.6. The base of the cone is depicted as being flat for the sake of clarity of the picture, but the reader should rather think of it as a part of a sphere.)

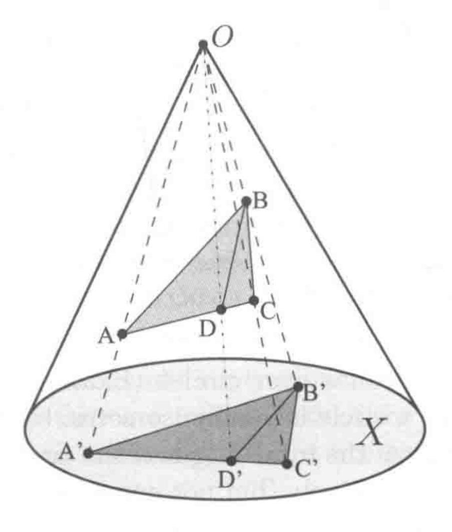


Figure 4.6: Cone is a space of $K \geq 0$ if and only if the space itself has $K \geq 1$.

We are going to prove that $\triangle abc$ in K satisfies the triangle condition for curvature ≥ 0 or ≤ 0 if and only if $\triangle a'b'c'$ in X does so for curvature ≥ 1 or ≤ 1, respectively, provided that the perimeter of $\triangle a'b'c'$ is no greater than 2π.

Consider a comparison triangle $\triangle\bar{a}'\bar{b}'\bar{c}'$ for $\triangle a'b'c'$ in the standard unit sphere $S^2 \subset \mathbb{R}^3$ (centered at the origin $\overline{O} \in \mathbb{R}^3$). Place points $\bar{a}, \bar{b}, \bar{c}$ in the rays $\overline{O}\bar{a}', \overline{O}\bar{b}', \overline{O}\bar{c}' \subset \mathbb{R}^3$ respectively so that $|\overline{O}\bar{a}| = |Oa|$, $|\overline{O}\bar{b}| = |Ob|$ and $|\overline{O}\bar{c}| = |Oc|$. The resulting triangle $\triangle\bar{a}\bar{b}\bar{c}$ in $\mathbb{R}^3$ has the same side lengths as $\triangle abc$; in other words, $\triangle\bar{a}\bar{b}\bar{c}$ is a comparison triangle for $\triangle abc$.

Pick a point $d \in [ac] \subset K$ and let $d' \in [a'c'] \subset X$ be the projection of d. Let $\bar{d}$ and $\bar{d}'$ be the corresponding points in the Euclidean segment $[\bar{a}\bar{c}]$ and the spherical segment $[\bar{a}'\bar{d}']$, respectively. The sub-cone over $[a'c']$ in K is isometric to the planar sector in $\mathbb{R}^3$ spanned by the spherical segment $[\bar{a}'\bar{c}']$. An isometry from this sub-cone to this sector sends $[ac]$ isometrically to the segment $[\bar{a}\bar{c}]$; in particular, it sends d to $\bar{d}$. Furthermore, this isometry is naturally related to the isometry from $[a'c']$ to the spherical segment $[\bar{a}'\bar{c}']$, and the latter sends d' to $\bar{d}'$. It follows that $\bar{d}$ belongs to the ray $\overline{O}\bar{d}'$ and $|\overline{O}\bar{d}| = |Od|$.

Assuming the distances $|Ob|$ and $|Od|$ fixed, the distance $|bd|$ in K is an increasing function in $|b'd'|$ (recall the formula for the distance in a cone). If $|b'd'| = |\bar{b}'\bar{d}'|$, then $|bd| = |\bar{b}\bar{d}|$ because $|\overline{O}\bar{b}| = |Ob|$ and $|\overline{O}\bar{d}| = |Od|$. Hence $|b'd'| \geq |\bar{b}'\bar{d}'|$ if and only if $|bd| \geq |\bar{b}\bar{d}|$, and vice versa. This proves the desired statement about distance conditions for $\triangle abc$ and $\triangle a'b'c'$.

2. To finish the proof of the theorem, it remains to connect the property proved in the first step to the definitions, and consider various marginal cases.

First, observe that every triangle with side lengths less than π is a projection of some triangle in K (which can be placed in an arbitrarily small neighborhood of the origin). Thus, by the result of the first step, if K has curvature ≥ 0 (resp. ≤ 0), then X has curvature ≥ 1 (resp. ≤ 1) in the large.

Similarly, the projection of any normal region in $K \setminus \{O\}$ (for curvature ≥ 0 or ≤ 0) is a normal region in X (for curvature ≥ 1 or ≤ 1 respectively), and, conversely, a sub-cone over a (sufficiently small) normal region in X is a normal region in $K \setminus \{O\}$. ("Sufficiently small" means, for example, of diameter less than $2\pi/3$.) This proves the first statement of the theorem.

3. Now let us consider "large" triangles in X. Namely, let $\triangle abc$ be a triangle in K whose sides do not pass through O but whose projection to X, $\triangle a'b'c'$, has perimeter $L > 2\pi$. Then the sub-cone over $\triangle a'b'c'$ in K (composed of three convex flat planar sectors) is an image under an arcwise

isometry of the cone K_1 over the circle of length L. Moreover, $\triangle abc$ is an image of a triangle in K_1. Since an arcwise isometry is a nonexpanding map and K_1 has nonpositive curvature, the triangle $\triangle abc$ satisfies the triangle condition for curvature ≤ 0 (without any conditions on X).

On the other hand, observe that the triangle in K_1 which corresponds to $\triangle abc$ does not satisfy the triangle condition for curvature ≥ 0 because the origin of K_1 is "inside" this triangle. Neither does the triangle $\triangle abc$. It follows that, if K has curvature ≥ 0, then X contains no triangles with sides less than π and perimeter greater than 2π.

Triangles $\triangle a'b'c'$ in X with some of sides greater than or equal to π are considered similarly. Such a triangle corresponds to a triangle in K some of whose sides pass through O, and the latter is an image under an arcwise isometry of a triangle in a (nonpositively but not nonnegatively curved) cone over a segment of length $L > \pi$. (Compare with the last part of the proof of Theorem 4.2.14 about polyhedral spaces.)

The conclusion is: for a connected X, existence of triangles with perimeter greater than 2π does not affect nonpositive curvature of K but prevents K from being nonnegatively curved. This proves the theorem in the case when X is connected.

4. Finally, consider the case when X is not connected. Then X is a disjoint union of its components, and K is a metric bouquet of cones over the components. By Proposition 4.2.9, K has curvature ≤ 0 if and only if the cone over each component has curvature ≤ 0. This finishes the proof of the second statement of the theorem.

And K cannot have curvature ≥ 0 unless X is a two-point space (in which case $K \cong \mathbb{R}$). Indeed, let x and y be two points in different components of X, and let $z \in X$ be an arbitrary third point. We may assume that x and z belong to different components (otherwise interchange x and y). Let x', y' and z' be the corresponding points in K; namely, $x' = (x, 1)$, $y' = (y, 1)$ and $z' = (z, 1)$. Then the sides $[x'y']$ and $[x'z']$ of $\triangle x'y'z'$ pass through O and share a common segment $[x'O]$. Hence $\measuredangle y'x'z' = 0$, so $\triangle x'y'z'$ does not satisfy the angle condition.

Gathering all implications proved during this long argument, one can easily see that the theorem follows. $\square$

Chapter 5

Smooth Length Structures

This chapter deals with certain length spaces whose definition and analysis involve smooth structures (calculus of variations). We begin with a basic discussion of Finsler and Riemannian manifolds, followed by a metric introduction to hyperbolic geometry. Then we consider some other interesting examples. Here is a brief plan of the chapter along with some guidance through it.

Section 5.1 describes Riemannian metrics in coordinates and promotes an important concept that a Riemannian metric locally is almost Euclidean. For the sake of simplicity of notations we mainly stick to two-dimensional regions; generalizations to higher-dimensional manifolds are obvious. In Section 5.2 we exploit an observation that shortest paths in a Riemannian metric satisfy a particular second-order differential equation. This equation enables us to draw geodesics with given initial positions and velocity vectors.

Using this, we can build *normal coordinate systems*. They are analogous to polar, spherical, and Cartesian coordinates: one family of coordinate lines is formed by geodesics parameterized by arc length, and the other family consists of equidistant curves orthogonal to the curves from the first family.

There are two ways of regarding normal coordinates. These two approaches correspond to two dual viewpoints in mechanics and optics: propagation of rays and wavefronts. Namely, one can first begin with a family of geodesics (for instances, emanating from one point—thus obtaining normal coordinates centered at the point). These geodesics can be perceived as rays, and their orthogonal trajectories, which happen to be equidistant

curves, correspond to wavefronts. Alternatively, one can start with a smooth curve (initial wavefront), and draw a family of its equidistants—curves lying at a fixed distance from the initial wavefront. This family represents propagation of the initial wavefront, and orthogonal trajectories to the family of wavefronts happen to be geodesics. To generalize the notion of normal coordinates to higher dimensions, one considers a family of geodesics forming orthogonal trajectories for a family of equidistant hypersurfaces.

There is no essential difference between Finsler and Riemannian metrics till here, and results of the first two sections hold (with appropriate modifications) for Finsler manifolds as well.

There is a remarkable one-parameter family of fully homogeneous two-dimensional Riemannian metrics formed by the spheres, the Euclidean space and the hyperbolic spaces. These are two-dimensional metrics of constant curvature, and their geometry is relatively well understood. Geometric meaning of the Gaussian curvature transpires via comparison with these spaces. Section 5.3 gives a metric introduction to two-dimensional hyperbolic geometry. Hyperbolic geometry can be regarded from an axiomatic viewpoint, as well as by means of Riemannian geometry. Its study is, however, assisted a lot by exploiting an auxiliary Euclidean structure: we use a model of the hyperbolic plane based on Euclidean geometry.

Section 5.4 introduces a remarkable example of sub-Riemannian length structures. Whereas Riemannian metrics are obtained by modifying the Euclidean length functional, sub-Riemannian length structures are defined by restricting the class of admissible curves. We limit ourselves to presenting some distinctive features of a model example of sub-Riemannian metrics. This section is optional, and it assumes some knowledge of Lie bracket and differential forms.

Two last sections of the chapter deal with volumes in smooth length spaces.

5.1. Riemannian Length Structures

5.1.1. Tangent vectors and coordinate systems. There is a good tradition that almost every textbook in geometry begins with an introduction to the theory of smooth manifolds. We will try to minimize this introduction, reducing it to notation and terminology accompanied by a brief ideological discussion. All we need is to introduce the space $T\Omega$ of vectors tangent to a planar region Ω. For a reader who finds it difficult to get used to notions introduced in this section, we advise skipping it for the first time and just thinking of $T\Omega$ as $\Omega \times \mathbb{R}^2$, with a Cartesian coordinate system (x, y) in Ω, and two base vectors $X = (1,0)$, $Y = (0,1)$ in the second factor $\mathbb{R}^2$. This

noninvariant coordinate viewpoint is quite convenient as long as we do not change coordinates.

Tangent vectors. For a point $p \in \Omega$, consider all smooth curves $\gamma\colon [0,1] \to \Omega$ starting from p, that is, with $\gamma(0) = p$. Two curves are said to be equivalent if their derivatives (velocities) at 0 are the same, that is, if they start from p with the same velocity. A *tangent* vector v at p is an equivalence class of such curves $v = [\gamma]$.

Of course, since all equivalent curves start from p with the same velocity vector, it appears as that there is no essential difference between tangent vectors and usual vectors. Hence a reader may wonder why we make an easy notion difficult. The reason for that becomes clear when one wants to change a coordinate system. Indeed, though we stated that Ω is a planar region, we do not want to fix such an identification. For instance, if Ω is a surface in $\mathbb{R}^3$, it can be identified with a region in $\mathbb{R}^2$, but there are lots of such identifications (coordinate systems), and possibly no preferred one. Secondly, even if Ω is already a subset of $\mathbb{R}^2$, re-imbedding it in a more convenient way may still be helpful.

As a matter of fact, by introducing tangent vectors we rather make a difficult object easy. When studying high school geometry, it takes a serious effort to get used to the concept that two vectors applied at different points "are equal" if their magnitudes and directions are. The main message now is that tangent vectors at different points *cannot be equal*, nor they can be added.

Even a consistent concept of "the same direction at different points" cannot exist in general—because otherwise one could easily produce a non-vanishing vector field tangent to a sphere (and this is a classical theorem that it is impossible to "comb the hair on a ball").

The set of tangent vectors at p will be denoted by $T_p\Omega$; the set of all tangent vectors will be denoted by $T\Omega$. Since tangent vectors at p can be identified with usual vectors, one can add them and multiply them by scalars. These operations are invariant, that is, the result is independent of the coordinate system. Thus $T_p\Omega$ is a (two-dimensional) vector space. Although tangent vectors at different points formally can be added as well (if we identify them with Euclidean vectors), we will not do this. The reason is again that this "addition" does not persist under coordinate changes.

Let $\varphi\colon \Omega \to \Omega'$ be a smooth map. For a curve γ starting at p, its image $\varphi \circ \gamma$ starts at $\varphi(p)$; moreover, if two curves emanate from x at the same velocity, so do their images from $\varphi(p)$. Hence φ induces a map $d_p\varphi\colon T_x\Omega \to T_{\varphi(p)}\Omega'$, and more generally a map $d\varphi\colon T\Omega \to T\Omega'$. This map

is called the *derivative* of φ. One can check that it is a linear map on each $T_p\Omega$.

Coordinate systems. By a coordinate system we understand a diffeomorphism $\varphi\colon \Omega' \to \Omega$ from a region $\Omega' \subset \mathbb{R}^2$ in the coordinate Euclidean plane to Ω. For two reals x, y (with $(x,y) \in \Omega'$) we say that the point $\varphi(x,y) \in \Omega$ is the point with coordinates (x,y).

When a coordinate system is fixed, we will usually abuse notation and denote a point $\varphi(x,y)$ by (x,y), thus identifying Ω and Ω'!

Once we fix a coordinate system in Ω, we get the corresponding coordinate system in $T\Omega$. Indeed, for a point $p = (x,y)$ in Ω, there are two designated curves starting from p: $\gamma_x(t) = (x+t, y)$ and $\gamma_y(t) = (x, y+t)$. These curves are called coordinate lines (the x-line and the y-line) passing through p, and the corresponding tangent vectors are called the coordinate vectors. We denote them by X and Y (or by $X(p)$ and $Y(p)$ if we want to emphasize their dependence on p).

Recall that a vector field is a correspondence (a map) that assigns for every point a tangent vector at this point. Thus the coordinate vectors form two coordinate vector fields.

If our coordinate system is given by a map φ, the coordinate vectors X and Y are just the images of the vectors of the standard basis $\mathbf{i} = (1,0), \mathbf{j} = (0,1) \in \mathbb{R}^2$ under $d\varphi$. The coordinate vector fields are also often denoted by $\partial/\partial x$ and $\partial/\partial y$. The reason for the latter notation becomes obvious in the next paragraph. Now for a vector $V \in T_p\Omega$ we define its coordinates (v_x, v_y) to be such (uniquely defined) reals that $V = v_x X + v_y Y$. We will often write $V = (v_x, v_y)$, identifying again an object and its coordinates.

What are tangent vectors good for? The most important usage of a vector is to differentiate functions: if $f\colon \Omega \to R$ is a smooth function and $V = [\gamma]$ is a tangent vector at p, we define its derivative Vf along V to be

$$Vf = \left.\frac{df(\gamma(t))}{dt}\right|_{t=0}.$$

Differentiation with respect to the coordinate vector fields X and Y is nothing but taking the well-familiar partial derivatives with respect to x and y, thus explaining the notation $X = \frac{\partial}{\partial x}$, $Y = \frac{\partial}{\partial y}$. Moreover, tangent vectors can even be defined axiomatically as "derivations", that are maps from the space of the smooth functions to reals satisfying certain properties (such as linearity over scalars, and the product rule).

Another way of using vectors is to "move with the velocity prescribed by a vector field". More formally, for a (sufficiently smooth) vector field V, one can consider its integral curves. An integral curve is a curve $\gamma(t)$ such

that $\gamma'(t) = d\gamma(t)/dt = V(\gamma(t))$ for all t. The existence and uniqueness of an integral curve with a fixed initial condition ($\gamma(0) = p$) is a principal theorem in the theory of ordinary differential equations.

Degenerate coordinates. It is convenient to consider more general "degenerate" coordinate systems. For instance, polar coordinates (r, ρ) are degenerate at $r = 0$: the map $(r, \rho) \longrightarrow (x = r\cos\rho,\, y = r\sin\rho)$ is not invertible, and its derivative is degenerate at $r = 0$. By a degenerate coordinate system we understand a smooth surjective map $\varphi\colon \Omega' \subset \mathbb{R}^2 \to \Omega$. We assume that the degenerations of φ are nice, as in polar coordinates: they take place on a few smooth connected submanifolds, each of which is mapped to a point, and φ is a local diffeomorphism elsewhere. In many cases, such as polar or normal coordinates, a "degenerate coordinate system" is traditionally called a coordinate system; we will follow this abuse of terminology. One can work with degenerate coordinate systems in almost the same way as with usual coordinate systems. The only differences are that there may be a few points in U that have many pre-images, and that the coordinate vector fields should be regarded as vector-valued maps $X, Y : \Omega' \to T\Omega$ given by

$$X(x, y) = d\varphi_{(x,y)} \frac{\partial}{\partial x} \in T_{\varphi(x,y)}\Omega,$$
$$Y(x, y) = d\varphi_{(x,y)} \frac{\partial}{\partial y} \in T_{\varphi(x,y)}\Omega.$$

This means that one may have to assign multiple values of a coordinate vector field at a degenerate point. For instance, $\frac{\partial}{\partial r}$ looks like a porcupine at the origin of polar coordinates: it consists of unit vectors sticking in all directions (on the other hand, $\frac{\partial}{\partial \rho}$ is zero at the origin). Notice that if two vector fields V, W are coordinate vector field for a coordinate system, this in particular means that they are linearly independent at every point. This restriction is no more in effect for two vector fields arising from a degenerate coordinate system. Nevertheless, only a special class of pairs of vector fields can be obtained this way. These are "commuting vector fields", which will be discussed later.

5.1.2. Finsler surfaces. Let us recall that Finsler length structures, as well as their particular case of Riemannian and conformal length structures, are motivated by the idea that we think of the distance as the time (or other resources) needed to travel between two points. We start from an intuitive assumption that the rate per "unit of Euclidean distance" depends on where we travel and in what direction. This is formalized by a smooth nonnegative "cost" function $\lambda\colon T\Omega \to \mathbb{R}$, where $\Omega \in \mathbb{R}^2$ is a region. We think of $(p, V) \in T\Omega$ as a point p and a "velocity vector" V at p. We can

think of $\lambda(p, V) = \lambda_p(V)$ as having two arguments representing location and velocity (we would love to denote this function by L, as this is nothing but the Lagrangian in mechanics; alas, L is already used to denote the length functional). Finslerian length structure generated by λ is given by all piecewise smooth curves together with a length functional defined by

$$L(\gamma, a, b) = \int_a^b \lambda_{\gamma(t)}(\gamma'(t))dt. \tag{5.1}$$

Recall that we impose an important requirement on λ: for each p, the function $\lambda_p(V) = \lambda(p, V)$ is a norm on $T_p\Omega$. Its positive homogeneity guarantees that L is invariant under reparameterizations of curves, and its convexity (the triangle inequality) implies the semi-continuity of L.

Exercise 5.1.1. Show that the Finsler length structure defined by (5.1) is indeed a lower semi-continuous length structure, and therefore it is induced by its intrinsic metric.

As we are going to do local analysis, we assume that all considerations take place in a neighborhood that is remote enough from the boundary of Ω; for instance, we may assume that $\Omega = \mathbb{R}^2$. This assumption is made to stay away from boundary effects, such as the situation when a shortest curve "bends around a peninsula of the complement of the region".

As usual, the length structure L gives rise to an intrinsic metric d. It is easy to show that the length of a curve induced by d coincides with L for all smooth curves (Exercise 5.1.1). However, whereas the existence of a shortest path for d connecting any two points follows from Theorem 2.5.23, its smoothness is less obvious. Moreover, a shortest path for d can fail to be smooth unless a certain additional assumption of *strict convexity* is imposed on λ_p.

Exercise 5.1.2. To see how a shortest curve can be nonsmooth, consider a function λ_p which is given by the following norm (independent of p): $\lambda_p(v_1, v_2) = |v_1| + |v_2|$. Describe all shortest paths in the corresponding metric space. In particular, notice that every broken line consisting of two segments with end-points $(0, 0)$, $(1, 1)$ and an intermediate vertex (x, y), where $x \geq 0$, $y \geq 0$, $x + y \leq 1$, is a shortest path.

Such a pathology (nonsmoothness and nonuniqueness of shortest paths, even in an arbitrary small neighborhood of a point) is due to the fact that balls of the norm (which are "diamonds") are not strictly convex. An interested reader can prove that for strictly convex norms all shortest paths are smooth, and a shortest path connecting two sufficiently close points is unique.

Hint: Use an analog of Lemma 5.1.13 with an appropriate two-dimensional normed space instead of Euclidean space.

Now we are going to leave general Finsler metrics and concentrate on their special case of Riemannian metrics. Whereas the geometry of Finsler spaces may prove to be even more interesting than Riemannian geometry, it is now not nearly as developed. This is a rapidly growing field, its basic methods have not been settled yet, and it is not all clear what kind of problems will determine the progress of Finsler geometry even in the near future.

5.1.3. Riemannian metrics.

Metric coefficients. For Riemannian metrics, the functions λ_p are given by positive definite quadratic forms:

$$\lambda_p(V) = \sqrt{Q_p(V,\, V)},$$

where, for each p, Q_p is a positive definite quadratic form. We will use the notation $\langle V,\, W\rangle_p = Q_p(V, W)$ for the symmetrical bilinear form corresponding to Q_p. The quantity $|V| = \sqrt{\langle V,\, V\rangle} = \sqrt{Q(V, V)}$ is called the *length of a tangent vector* V.

Let us see how a Riemannian metric can be described using a coordinate system (x, y). To fix a quadratic form Q_p on $T_p\Omega$, it is enough to know the values

$$\begin{aligned}
E(p) &= Q_p(X(p),\, X(p)) = \langle X(p),\, X(p)\rangle\,,\\
F(p) &= Q_p(X(p),\, Y(p)) = Q_p(Y(p),\, X(p)) = \langle X(p),\, Y(p)\rangle\,,\\
G(p) &= Q_p(Y(p), Y(p)) = \langle Y(p),\, Y(p)\rangle\,,
\end{aligned}$$

which are the entries of the matrix of the bilinear form $\langle , \rangle_p$ in the basis formed by the coordinate vectors $X(p)$ and $Y(p)$. Indeed,

$$\begin{aligned}
(5.2)\quad Q(V, W) &= Q(v_xX + v_yY,\, w_xX + w_yY)\\
&= Ev_xw_x + Fv_xw_y + Fv_yw_x + Gv_yw_y
\end{aligned}$$

(to avoid cumbersome notation, we drop the dependence on p in this and many other formulas). Hence a Riemannian metric on Ω can be given by its *metric coefficients*, which are three functions E, F, G: $\Omega \to \mathbb{R}$. These functions have to satisfy the inequalities $E > 0$, $G > 0$, $EG - F^2 > 0$ since Q is a positive definite quadratic form. A reader who is not familiar with quadratic forms can take the formula (5.2) as a coordinate definition of a Riemannian metric. Of course, coordinate vectors of a degenerate coordinate system may be linearly dependent; one can still define $E = \langle X,\, X\rangle$, $F = \langle X,\, Y\rangle$ and $G = \langle Y,\, Y\rangle$, but it is possible to claim only that $E \geq 0$, $G \geq 0$, $EG - F^2 \geq 0$.

Exercise 5.1.3. Show that if one introduces new coordinates (u, v) such that $x = x(u, v)$, $y = y(u, v)$ the rule of converting E to the new coordinates is given by the following formula:

$$E_{new} = E(\frac{\partial x}{\partial u})^2 + 2F\frac{\partial x}{\partial u}\frac{\partial y}{\partial u} + G(\frac{\partial y}{\partial u})^2.$$

Find analogous formulas for F and G.

Note: Recall that a coordinate system is formally a map $\varphi\colon \Omega_1 \to \Omega$. Changing coordinates to (u, v) such that $x = x(u, v)$, $y = y(u, v)$, where $x(u, v)$ and $y(u, v)$ are two given functions (it is just convenient to denote them by the same letters as coordinates) formally means that we consider a map $\psi\colon \Omega_2 \to \Omega_1$, $\psi(u, v) = (x(u, v), y(u, v))$, and our new coordinate system is a map $\varphi \circ \psi\colon \Omega_2 \to \Omega$.

Once a coordinate system (x, y) is chosen, the length of a curve $\gamma(t) = (x(t), y(t))$ is given by

$$L(\gamma, a, b) = \int_a^b \sqrt{E(x')^2(t) + 2Fx'(t)y'(t) + G(y')^2(t)}\, dt \tag{5.3}$$

where $E = E(x(t), y(t))$ and similarly for F and G.

5.1.4. Isometries and Riemannian manifolds. Let us give the following (obvious) definition:

Definition 5.1.4. Two regions Ω, Ω' with Riemannian metrics Q, Q' are said to be *isometric* if there exists a diffeomorphism $\varphi\colon \Omega \to \Omega'$ such that $Q'(d\varphi(V), d\varphi(V)) = Q(V, V)$ for every tangent vector $V \in T\Omega$.

Equivalently, two Riemannian metrics are isometric if there exist coordinate systems in Ω and Ω' such that the metric coefficients E, F, G for the metrics are the same at points with the same coordinates.

Remark 5.1.5. Isometric Riemannian regions are obviously isometric as length spaces. The converse is also true: if Riemannian regions Ω and Ω' are isometric as length spaces, then they are isometric in the sense of Definition 5.1.4 (that is, there exists a *smooth* isometry that respects their Riemannian structures). Moreover, *every* isometry map from Ω to Ω' is smooth, and (as a consequence) can be taken as φ in Definition 5.1.4.

We leave this fact as an (not so obvious!) exercise. The easiest proof we know is based on the results of Section 5.2 (namely, smoothness of shortest paths and properties of exponential maps).

Smoothness of isometries allows us to give a *metric* definition of a Riemannian manifold (for we were so far looking only at regions with Riemannian metrics):

Definition 5.1.6. A *Riemannian manifold* is a length space such that every point has a neighborhood isometric to a region with Riemannian metric.

Remark 5.1.7. This definition is not standard. In most textbooks Riemannian manifolds are defined as *smooth manifolds* equipped with Riemannian structures. The definitions are equivalent: this easily follows from Remark 5.1.5 (in particular, a length space that is locally isometric to a Riemannian region naturally carries a structure of a smooth manifold). If you are familiar with smooth manifolds, we recommend you prove this equivalence as an exercise.

If you are not happy with these abstract definitions, you may think of a Riemannian manifold as a smooth surface in a Euclidean space. With this simplification, one does not even sacrifice generality: according to the famous Nash's Embedding Theorem [**N**], every Riemannian manifold is isometric to a smooth embedded surface in a Euclidean space of some (sufficiently large) dimension. Two-dimensional examples of this kind are discussed in the next subsection.

Embedded surfaces. An important example of Riemannian metrics comes from embedded surfaces. This is one of the main motivating examples for the theory; it goes back to K. F. Gauss. If $r\colon \Omega \to \mathbb{R}^3$ is an embedding, one defines $\lambda_p(V) = |d_p r(V)|$. The corresponding bilinear form is given by the pull-back of the Euclidean scalar product under the embedding: $\langle V, W\rangle_p = \langle d_p r(V), d_p r(W)\rangle_E$, where $\langle , \rangle_E$ is the Euclidean scalar product in $\mathbb{R}^3$. This definition has a very clear geometric meaning: if γ is a smooth curve in Ω, the length of γ is equal to the Euclidean length L_E of its image under r. Indeed,

$$\begin{aligned} L(\gamma, a, b) &= \int_a^b \lambda_{\gamma(t)}(\gamma'(t))\,dt \\ &= \int_a^b |d_{\gamma(t)} r(\gamma'(t))|\,dt = \int_a^b \big|\frac{d}{dt} r(\gamma(t))\big|\,dt = L_E(r \circ \gamma, a, b). \end{aligned}$$

A reader may very well think of Ω as surface in $\mathbb{R}^3$ and regard r as a coordinate system; then dr disappears from the formula.

This computation is valid only for smooth curves. Notice that we have Ω with its Riemannian metric, and we have its image in $\mathbb{R}^3$ with the *intrinsic metric d_I induced from the ambient Euclidean distance.* We want to show that r is an isometry between the two metric spaces. In other words, we introduce a metric on $r(\Omega)$ via the class of all piece-wise smooth curves. Of course, r is an isometry with respect to this metric, so we will also denote it by d (this would not even be an abuse of notation if we think of Ω as already a subset of $\mathbb{R}^3$ with a coordinate system given by r). We want to

see that $d = d_I$. There is an obvious inequality $d \geq d_I$. The other inequality is contained in the following exercise:

Exercise 5.1.8. 1. For two (sufficiently close) points $p, q \in r(\Omega)$, let $\rho(p,q)$ be the length of the curve $\gamma_{pq} = r(\Omega) \cap P$, where P is the plane containing the segment $[pq]$ and the normal to the surface $r(\Omega)$ at p (notice that γ_{pq} is actually a smooth curve provided that q is sufficiently close to p). Show that

$$\rho(p,q) - |pq| = o(|pq|)$$

as $|pq| \to 0$, where $|pq|$ is Euclidean distance in $\mathbb{R}^3$.

2. Show that, given a curve of length L in $r(\Omega)$ and $\varepsilon > 0$, there is a piece-wise smooth curve with the same endpoints and whose length is at most $L - \varepsilon$.

Hint: Use directly the definition of an induced metric to find the induced length of a nonsmooth curve; in this definition, replace each segment $[p_i p_{i+1}]$ of a Euclidean broken line inscribed in the curve by a smooth curve γ_{pq}, and use the first part of the exercise to estimate how this could increase the length of the broken line.

It is still not clear yet that the shortest paths are smooth. This issue is discussed in the next section. We finish this section with a brief comment on the relationship between intrinsic and extrinsic geometries of a surface.

By *intrinsic geometry* we understand the properties that depend only on the Riemannian metric. In case of an embedded surface the Riemannian metric is induced by an embedding r.

It is easy to give examples of different embeddings that induce the same Riemannian structure.

Exercise 5.1.9. Verify that the embeddings $\varphi, \psi\colon [0,1] \times [0,1] \to \mathbb{R}^3$ given by

$$\varphi(u,v) = (u, v, 0),$$
$$\psi(u,v) = (\sin u,\ \cos u,\ v)$$

induce the same Riemannian metric.

A more interesting example is given in the next exercise. It shows that there is a whole variety of surfaces whose Riemannian metrics are isometric to those of planar regions:

Exercise 5.1.10. Let $\gamma\colon \mathbb{R} \to \mathbb{R}^3$ be a curve parameterized by arc length, and assume that the curvature of γ does not vanish. Consider an embedding $\varphi\colon U \Subset \mathbb{R}^2 \to \mathbb{R}^3$ from a sufficiently small neighborhood of a point $(u_0,\ v_0)$ with $v_0 \neq 0$ given by

$$\varphi(u,v) = \gamma(u) + v\dot\gamma(u).$$

Show that U with the Riemannian metric induced by this embedding is isometric to a flat region.

Hint: Consider an analogous embedding for a planar curve γ_1 with the same curvature (as a function of its natural parameter) as that of γ. It is clear that this embedding sends U to a planar region; show that it induces the same Riemannian metric as φ.

To show that two Riemannian metrics are *not* isometric, one looks for a metric property that can tell them apart. Here is an example:

Exercise 5.1.11. Give an elementary argument showing that a region of a sphere is not isometric to a planar region. More precisely, let $r\colon \Omega \to \mathbb{R}^3$ be an embedding whose image belongs to a sphere (for instance, $r(\varphi,\rho) = (\sin\varphi\sin\rho, \sin\varphi\cos\rho, \cos\varphi)$ defined on a small neighborhood in the (φ,ρ)-plane not intersecting the φ-axis). Prove that Ω with the Riemannian metric induced by r is not isometric to a Euclidean region.

Hint: Perhaps the easiest metric invariant that distinguishes the spherical regions from the Euclidean ones is the length of a circle (as a function depending on the radius).

This is quite a delicate problem to determine what properties of an embedding (extrinsic properties) can be determined by the intrinsic geometry of a surface. The most famous statement of this type is a theorem of K. F. Gauss (which delighted him so much that he called it "Theorema Egregium", that is, "Magnificent Theorem") stating that the product of principal curvatures of an embedded surface is an invariant of its intrinsic geometry. This invariant is called the *Gaussian curvature*, and in our exposition it will be introduced in terms of the intrinsic geometry. In particular, "Theorema Egregium" implies that the product of two principal curvatures does not change under *bendings* of a surface (a bending of a surface is a continuous family of embeddings inducing the same lengths of curves in the surface).

5.1.5. Riemannian: infinitesimally Euclidean. This section is devoted to a technically useful (and ideologically very important) observation that locally a Riemannian metric is almost Euclidean.

Choosing a coordinate system means choosing an identification between Ω and a planar region. Hence every coordinate system induces an auxiliary Euclidean structure in Ω. We are going to show that the coordinates can be chosen in such a way that the auxiliary Euclidean metric induced by the choice of the coordinates is close to the Riemannian one.

We will need the following lemma, whose proof is left to the reader (as a trivial exercise; this is just a linear coordinate change):

Lemma 5.1.12. *Given a point p, one can choose coordinates such that the metric coefficients at p are $E = G = 1$, $F = 0$. More precisely, one can choose a coordinate system $\varphi\colon \Omega' \subset \mathbb{R}^2 \to \Omega$, $\varphi(x_0, y_0) = p$ such that the metric coefficients at p are $E(x_0, y_0) = G(x_0, y_0) = 1$, $F(x_0, y_0) = 0$ in this coordinate system.*

Thus one can choose a coordinate system such that, at one given point p, the metric coefficients look the same as the Euclidean metric in Cartesian coordinates. This means that the coordinate vectors form an orthonormal basis of $(T_p\Omega, \langle,\rangle_p)$. In other words, the derivative $d_{(x_0,y_0)}\varphi$ of the coordinate map is a linear isometry

$$d_{(x_0,y_0)}\varphi\colon (T_{(x_0,y_0)}\mathbb{R}^3, \text{ Euclidean Scalar Product}) \to (T_p\Omega, \langle,\rangle_p).$$

The following lemma shows that if we identify Ω with a Euclidean region by a coordinate system as in Lemma 5.1.12, then the Riemannian and Euclidean distance functions are close to each other in a small neighborhood of p. Intuitively, this means that locally a Riemannian metric is "almost Euclidean":

Lemma 5.1.13. *Let a Riemannian metric on $\Omega \subset \mathbb{R}^2$ be such that its metric coefficients with respect to the standard Euclidean coordinates (x, y) in $\mathbb{R}^2$ are $E(p) = G(p) = 1$, $F(p) = 0$. Then, for every vector V at q, its Riemannian length $|V| = \sqrt{\langle V, V\rangle}$ is close to its Euclidean length $|V|_E$:*

$$\lim_{|pq|_E \to 0} \frac{|V|}{|V|_E} = 1.$$

Furthermore, the Riemannian distance function is close to the Euclidean one in a small neighborhood of p, namely,

$$\lim_{|pq|+|pr| \to 0} \frac{d(q, r)}{|qr|} = 1,$$

where $|pq|$ is the Euclidean distance between p and q.

The lemma follows from the continuity of the Riemannian metric Q, that is, from the continuity of its metric coefficients E, F, G. In its turn, the lemma easily implies that the Riemannian angle coincides with that in the length space induced by the Riemannian metric:

Lemma 5.1.14. *Let γ_1, γ_2 be two smooth paths in Ω starting from p with nonzero velocities, that is $\gamma_1(0) = \gamma_2(0) = p$ and $\gamma_1'(0) \neq 0$, $\gamma_2'(0) \neq 0$. Then the (metric) angle between γ_1 and γ_2 at p does exist and is equal to the Riemannian angle*

$$\arccos \frac{\langle \gamma_1', \gamma_2'\rangle}{\sqrt{\langle \gamma_1', \gamma_1'\rangle}\sqrt{\langle \gamma_2', \gamma_2'\rangle}}$$

between their velocity vectors $\gamma_1' = \gamma_1'(0)$ and $\gamma_2' = \gamma_2'(0)$ at $t = 0$.

Exercise 5.1.15. Prove Lemmas 5.1.13 and 5.1.14.

Remark. The arguments we know are absolutely straightforward but somewhat tedious.

There is the following convenient reformulation of Lemma 5.1.13. Let $\varphi\colon \Omega' \subset \mathbb{R}^2 \to \Omega$ be a coordinate system in Ω, and $p \in \Omega$. Let us introduce a scalar product $\langle , \rangle_P$ on $\mathbb{R}^2$ such that $d_p\varphi\colon (\mathbb{R}^2, \langle , \rangle_P) \to (T_p\Omega, \langle , \rangle_p)$ is an isometry. Recall that by definition the coordinate vectors X and Y at p are given by

$$X(p) = d_p\varphi(\frac{\partial}{\partial x}) = d_p\varphi(1,0), \quad Y(p) = d_p\varphi(\frac{\partial}{\partial y}) = d_p\varphi(0,1).$$

Hence in Euclidean coordinates the scalar product $\langle , \rangle_P$ is given by

$$\langle (1,0), (1,0)\rangle_P = E(p), \quad \langle (1,0), (0,1)\rangle_P = F(p), \quad \langle (0,1), (0,1)\rangle_P = G(p),$$

where E,F,G are the metric coefficients of the Riemannian metric.

Lemma 5.1.16. *For every vector V at q, its Riemannian length $|V| = \sqrt{\langle V, V\rangle}$ is close to its Euclidean length $|V|_P = \sqrt{\langle V, V\rangle_P}$ with respect to the Euclidean structure given by $\langle , \rangle_P$:*

$$\lim_{|pq|_P \to 0} \frac{|V|}{|V|_P} = 1.$$

Lemma 5.1.16 suggests the following way of computing (or defining) the Riemannian area in coordinates. Let us look at a coordinate rectangle $[x_0, x_0 + \Delta x] \times [y_0, y_0 + \Delta y]$, where $p = (x_0, y_0)$, and Δx, Δy are two (small) positive numbers. Then Lemma 5.1.16 suggests that a definition of Riemannian area should be such that the area of this rectangle is close to its Euclidean area with respect to $\langle , \rangle_P$. It is a standard exercise in Euclidean vector computations to show that the area of a parallelogram spanned by vectors V and W is equal to $\sqrt{\langle V, V\rangle \langle W, W\rangle - \langle V, W\rangle^2}$. Hence the Euclidean $\langle , \rangle_P$-area of the rectangle is equal to $\sqrt{E(p)G(p) - F^2(p)}\,\Delta x \Delta y$. Thus one defines the Riemannian area of Ω by integration as

$$\text{Area}(\Omega) = \int_{\Omega'} \sqrt{E((x,y))G(x,y) - F^2(x,y)}\, dxdy, \tag{5.4}$$

where Ω' is the corresponding region in the (x, y)-plane.

Exercise 5.1.17. 1. Prove Lemmas 5.1.16.

2. Show that the Riemannian area defined by (5.4) coincides with the 2-dimensional Hausdorff measure.

3. Generalize the formula (5.4) to higher dimensions.

Smoothness of shortest paths. This subsection is optional; we will discuss the smoothness of shortest paths, which will follow independently from further results as well. It is not difficult to prove that all naturally parameterized shortest paths in Ω are differentiable. Unfortunately, we know only a rather tedious, though a straightforward argument. We will not present a complete proof here, limiting ourselves to its punchline.

We want to prove that a curve γ is differentiable at t_0. Denote $\gamma(t_0) = p$.

Let us introduce a coordinate system in Ω, thus introducing an auxiliary Euclidean structure in Ω. Applying Lemma 5.1.12, we may assume without loss of generality that $E(p) = G(p) = 1$, $F(p) = 0$. There is a routine argument that shows that the differentiability of γ is equivalent to the fact that "γ has a certain direction and a certain speed", namely, that it satisfies the following two conditions:

$$\limsup_{u,v\to t, uv>0} \measuredangle(p,\gamma(u),\gamma(v)) = 0, \tag{5.5}$$

where $\measuredangle(A,B,C)$ is the Euclidean angle at A in the triangle ΔABC, and

$$\frac{d}{dt}\Big|_{t_0} |p\gamma(t)| = 1, \tag{5.6}$$

where $|pq|$ is the Euclidean distance.

Lemma 5.1.13 immediately implies Condition (5.6).

To check Condition (5.5), one considers a triangle $\Delta(p,\gamma(u),\gamma(v))$, where $u > v > 0$. It is a standard exercise in planimetry to check that, if the angle $\measuredangle(p,\gamma(u),\gamma(v)) \geq \alpha > 0$, there exists a positive constant $C = C(\alpha)$ such that

$$|p\gamma(u)| + |\gamma(v)\gamma(u)| - |p\gamma(v)| \geq C|p\gamma(v)|.$$

Combining this observation with the fact that all ratios

$$\frac{d(p,\gamma(u))}{|p\gamma(u)|}, \quad \frac{d(p,\gamma(v))}{|p\gamma(v)|}, \quad \frac{d(\gamma(u),\gamma(v))}{|\gamma(u)\gamma(u)|}$$

converge to 1 as $u, v \to 0$, one arrives at a contradiction to the fact that $d(p,\gamma(u)) = d(p,\gamma(v)) + d(\gamma(u),\gamma(v))$.

Notice that our further exposition will not rely on this argument; we will derive the smoothness of the shortest paths as an easy corollary of the Gauss Lemma 5.2.8.

5.1.6. Important examples. Let us explicitly compute the metric coefficients E, F, G for some "model" metrics in "natural" coordinates.

Let us begin with the Euclidean metric in polar coordinates (which degenerate at the origin). The coordinates are given by the map $(r,\rho) \longrightarrow (r\cos\rho,\ r\sin\rho)$, and hence the coordinate vectors are $\frac{\partial}{\partial r} = (\cos\rho, \sin\rho)$,

$\frac{\partial}{\partial\rho} = (-r\sin\rho,\, r\cos\rho)$ (where the right-hand side expressions are given in Cartesian coordinates). Hence

$$E(r,\rho) = 1,\ \ F(r,\rho) = 0,\ \ G(r,\rho) = r^2. \tag{5.7}$$

Notice that, although the Euclidean plane is absolutely "homogeneous", that is, it looks "the same" at every point, this is not immediately seen from (5.7): it is only obvious that the Euclidean plane is rotationally symmetric, for its metric coefficients do not depend on ρ, and hence a transformation $(r,\rho) \longrightarrow (r, \rho + \text{const})$ is an isometry.

Exercise 5.1.18. Represent a parallel translation in polar coordinates and verify that it is an isometry directly by checking that it preserves the length structure given by (5.7).

Now let us consider the sphere of radius R with the (degenerate) coordinates given by the map

$$(\varphi,\, \rho) \longrightarrow (x = R\sin(\varphi/R)\cos\rho,\ y = R\sin(\varphi/R)\sin\rho,\ z = R\cos(\varphi/R)).$$

This is almost the usual spherical coordinate system, and the only difference is that the φ-coordinate is rescaled to have the φ-lines parameterized by arc length. A trivial computation yields:

$$E = 1,\ \ F = 0,\ \ G = R^2\sin^2(\varphi/R). \tag{5.8}$$

Similarly to the previous case of the Euclidean plane, only the rotational symmetry $(\varphi,\, \rho) \longrightarrow (\varphi,\, \rho + \text{const})$ of the metric is obvious directly from the formulas. Of course, we know that a sphere is a perfectly symmetric space from Euclidean considerations: there is a rigid motion of $\mathbb{R}^3$ that maps the sphere to itself and sends any given point to any other given point. It is, however, not at all transparent from the intrinsic viewpoint when we look at the formula (5.8); if we *defined* a sphere as a surface whose Riemannian metric is given by (5.8), producing formulas for all isometries would be quite a task (compare with the previous exercise; a reader who likes spherical geometry may enjoy converting to spherical coordinates a 3-dimensional rotation about a line other than the z-axis).

Now we *define the hyperbolic plane of curvature* k (where $k < 0$!) as a plane with Riemannian metric given by the following formulas for its metric coefficients:

$$E = 1,\ \ F = 0,\ \ G = \frac{1}{-k}\sinh^2(\sqrt{-k}\, r). \tag{5.9}$$

Other definitions of the hyperbolic plane are given in Section 5.3. It will be shown later that the hyperbolic planes are as homogeneous as the spheres and the Euclidean plane.

Exercise 5.1.19. Using (5.8) and (5.9), compute the length of a circle of radius r in the sphere of a radius R and in the hyperbolic plane of a curvature k.

5.2. Exponential Map

The main objective of this section is to show that one can introduce a coordinate system in a neighborhood of every point such that E is identically 1 and F is identically 0. Such coordinate systems are called *normal coordinates.* It is easy to see that one family of coordinate lines in such coordinates has to be formed by shortest paths (parameterized by arc length). We introduce *geodesics* as curves satisfying a certain differential equation, and use them to build normal coordinates.

5.2.1. Geodesics. Normal coordinates. We define geodesics as the curves satisfying certain second-order differential equations. We will see that geodesics are locally shortest paths, and that every shortest path parameterized by arc length is a geodesic.

Definition 5.2.1. A *geodesic* is a smooth curve $(x(t), y(t))$ satisfying the following differential equations:

(5.10)
$$E\frac{d^2x}{dt^2} + F\frac{d^2y}{dt^2} = -\left((\frac{dx}{dt})^2\frac{\partial E}{\partial x} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial E}{\partial y} + (\frac{dy}{dt})^2\frac{\partial F}{\partial y} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial F}{\partial x}\right),$$

(5.11)
$$F\frac{d^2x}{dt^2} + G\frac{d^2y}{dt^2} = -\left((\frac{dx}{dt})^2\frac{\partial F}{\partial x} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial F}{\partial y} + (\frac{dy}{dt})^2\frac{\partial G}{\partial y} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial G}{\partial x}\right).$$

This definition may look strange at first glance. To motivate the choice of this differential equation, in the next chapter we will use variational methods to derive it as an equation that is satisfied by every smooth shortest path. A reader can notice that we actually *will not* use the result obtained by a variational argument. Instead, we will give a simple proof based on certain properties of the equation that the geodesics are the locally shortest curves. It will also become obvious later that this (coordinate) equation defines the same geometric object if we change the coordinate system.

The reader familiar with classical mechanics may also note that these equations describe a free particle whose kinetic energy is given by $E\dot{x}^2 + 2F\dot{x}\dot{y} + G\dot{y}^2$ (the quadratic form of the metric).

Exercise 5.2.2. Prove that every geodesic (i.e., every solution of equations (5.10) and (5.11)) except a constant map is a curve parameterized proportionally to arc-length.

Hint: Multiply equations (5.10) and (5.11) by $\frac{d}{dt}x$ and $\frac{d}{dt}y$, respectively, and then sum them up. After that compare the result with the equation

$$\frac{d}{dt}\left(E(\frac{dx}{dt})^2 + 2F\frac{dx}{dt}\frac{dx}{dt} + G(\frac{dxy}{dt})^2\right) = 0.$$

Exercise 5.2.3. Show by a direct computation that if a curve is a geodesic in a coordinate system, then it is a geodesic with respect to every coordinate system.

Hint: Use Exercise 5.1.3.

Now the main theorem of the theory of Ordinary Differential Equations tells us that, given an initial point $p = (x, y)$ and a unit vector $V = (v_x, v_y)$ at p, there exists a unique geodesic $\gamma(t)$ that starts from p at $t = 0$ with the velocity $\dot{\gamma}(0) = V$. (Prove that such a geodesic γ is really an arc-length parameterized curve.) The geodesic γ is defined for all sufficiently small $|t|$ (since the system is not linear, one cannot guarantee that such a solution is extendible to a bigger interval).

Definition 5.2.4. For a tangent vector $W \in T_p\Omega$, we define the exponential map $\exp_p(W)$ by

$$\exp_p(W) = \gamma(|W|),$$

where γ is the geodesic with $\gamma(0) = p$, $\dot{\gamma}(0) = W/|W|$, and $|W| = \sqrt{\langle W, W\rangle}$.

Geometric meaning of $\exp_p$ is very transparent: to arrive at $\exp_p(W)$, one travels along the geodesic starting from p in the direction of W for distance $|W|$.

Of course, $\exp_p(W)$ is not well-defined if $\gamma(t)$ cannot be extended for $t = |W|$, but at least it is defined for all sufficiently short vectors W. Hence we have constructed a map $\exp_p \colon \Theta_p \subset T_p\Omega \to \Omega$, defined on a neighborhood Θ_p of the origin in $T_p\Omega$. Regarding p as the second variable, we will also consider the map $\exp\colon \Theta \subset T\Omega \to \Omega$, $\Theta = \bigcup_{p\in\Omega} \Theta_p$. This is a smooth map (by the theorem that guarantees that the solution of the initial value problem smoothly depends on the initial condition).

Notice that the derivative of $\exp_p$ at $0 \in T_p\Omega$ is the identity map by the construction of $\exp_p$. Indeed, the image $\exp_p(tV)$ of a curve tV is a geodesic whose velocity vector at $t = 0$ is just V. (In particular, $\exp_p$ preserves the angles between rays emanating from the origin—a fact that we will extensively use later!) By the inverse function theorem $\exp_p$ is a diffeomorphism between a neighborhood U_p of the origin $0 \in T_p\Omega$ in the tangent space and its image $V_p = \exp_p(U_p)$ in Ω.

Injectivity radius. The maximum "size" of a neighborhood where $\exp_p$ remains a diffeomorphism is an important characteristic of the metric near p.

Definition 5.2.5. The injectivity radius r_p at p is the maximum real number r_p (or ∞) such that $\exp_p \colon B(r_p) \to \Omega$ is a diffeomorphism to its image, where $B(r)$ is the disc of radius r centered at $0 \in T_p\Omega$.

Exercise 5.2.6. Prove that $\inf\{r_p \colon p \in K\} \neq 0$ for any compact K contained in Ω.

Remark. This is a rather unpleasant exercise; though this is just some version of a compactness argument, a neat proof is rather involved. Perhaps the most straightforward way to do the exercise is to look at the proof of the implicit function theorem, and extract an effective estimate on the size of a neighborhood where the map is invertible. There are also nice geometrical proofs (using the sort of geometrical thinking that we are trying to promote by this book), but they are somewhat tricky.

Normal coordinates. Introduce a polar coordinate system $\varphi \colon \mathbb{R}^2 \to B(r_p)$ in the ball of the radius r_p centered at $0 \in T_p\Omega$. Combining it with $\exp_p$, we get a new coordinate system $(x, y) \to \exp_p \varphi(x, y)$ in $V_p = \exp_p B(r_p)$. To be more precise, this is a degenerate coordinate system, for polar coordinates are singular at the origin, but as usual this singularity does not cause any trouble. To find the coordinates of a point $q \in V_p \subset \Omega$, one connects p and q by a geodesic segment contained in V_p (its existence and uniqueness follow from the fact that $\exp_p \colon U_p \to V_p$ is a diffeomorphism). The length of this segment is the x-coordinate of q, and the angular coordinate of the velocity vector of this geodesic segment at p is the y-coordinate of q. This is just a Riemannian analog of polar coordinates. Such coordinates are called *normal coordinates centered at* p. Notice that the x-lines $\gamma_y(t) = (t, y)$ are geodesics parameterized by arc length.

More generally, *normal coordinates* is a coordinate system (x, y) such that its x-lines are geodesic parameterized by arc length, and the coordinate vector fields are orthogonal for $x = 0$. This means that the y-lines are geodesics orthogonal to the y-axis $x = 0$. In case of "generalized polar coordinates" described in the previous paragraph, the line $x = 0$ degenerates into one point p; to distinguish such "generalized polar coordinates" we called them *normal coordinates centered at* p.

The next section implies that alternatively normal coordinates can be defined as a coordinate system in which $E = 1$, $F = 0$.

Remark. The reader who follows a higher-dimensional version of the theory can define normal coordinates as a coordinate system such that the length of the first coordinate vector is 1, and it is orthogonal to the other coordinate vectors. This will imply that the coordinate lines corresponding to the first coordinate are geodesics. Intuitively it may be convenient to still think of two coordinates, the second one being $(n-1)$-dimensional; then one sees

a family of geodesic coordinate lines and a family of equidistant surfaces orthogonal to them. This approach is promoted in the next chapter (Section 6.4.1).

5.2.2. Gauss Lemma and local minimality of geodesics. There is a very simple criterion for a coordinate system to be normal:

Lemma 5.2.7. *If the metric coefficients with respect to a coordinate system satisfy $E = 1$, $F = 0$, then the coordinate system is normal.*

Proof. Equations (5.10) and (5.11) are satisfied by the x-lines ($x = t$, $y =$ const). Hence the lemma follows. □

We are going to show that the converse is also true, namely, that in normal coordinates the x-lines and the y-lines are mutually orthogonal (not only at $x = 0$). This fact is called the *Gauss Lemma*:

Lemma 5.2.8. *In normal coordinates the metric coefficient $F = \langle X, Y\rangle$ is identically zero.*

Proof. Consider a coordinate line $(x(t), y(t)) = (t, y_0)$. It has to satisfy equations (5.10) and (5.11) (note that we are using the result of Exercise 5.2.3). These equations for coordinate lines take a very simple form since $x' = 1$, $x'' = 0$, $y' = 0$, $y'' = 0$, where, as usually, prime mean derivative with respect to t. Thus we get two equations: $0 = -\partial F/\partial x$ and $0 = -\partial F/\partial x$. The first equation carries no new information. Indeed, the fact that the x-lines are parameterized by arc length is equivalent to $E \equiv 1$. The second equation implies that F does not change along the x-lines. Hence it is identically zero since it is 0 along the y-axis $x = 0$. □

By this lemma, in normal coordinates a Riemannian scalar product has only one nontrivial coefficient $G(x, y)$ since $E = 1$ and $F = 0$, and the coordinate lines form an orthogonal web.

We will employ this fact to show that a sufficiently short segment of a geodesic is indeed a shortest path:

Lemma 5.2.9. *Let $\gamma(t)$ be a geodesic with $\gamma(0) = p$, and let $b < r_p$, where r_p is the injectivity radius at p. Then γ is the shortest path between $p = \gamma(0)$ and $q = \gamma(b)$.*

Proof. Let us consider a piecewise smooth curve $\sigma(t) = (x(t), y(t))$, $t \in [0, b]$, connecting $p = (0, 0)$ and $q = (x_0, y_0)$, where (x, y) are normal coordinates centered at p. Notice that, by the construction of normal coordinates centered at p, $x_0 = b < r_p$. Let b_1 be the first value of t

such that $x(t) = x_0$. Then the segment $\sigma|_{[0,b_1]}$ is contained in our coordinate system, and we can estimate its length by integration:

$$L(\sigma) \geq \int_0^{b_1} \sqrt{(x')^2 + G(y')^2}\, dt \geq \int_0^{b_1} |x'|\, dt \geq \int_0^{b_1} x'\, dt = x_0.$$

Hence the length of any path from p to q is at least b. Analyzing the equality case in our inequalities, the reader can easily verify that it can happen only if $\sigma = \gamma$. □

Exercise 5.2.10. Show that a by-product of the argument above is the fact that all shortest paths are smooth!

Combining this lemma with Exercise 5.2.6 yields a very useful

Lemma 5.2.11. *Every point $p \in \Omega$ possesses a neighborhood such that every two points from this neighborhood can be connected by exactly one shortest path. Every geodesic segment contained in this neighborhood is the shortest path between its endpoints.*

Exercise 5.2.12. Prove the lemma.

5.3. Hyperbolic Plane

Hyperbolic planes are defined in Section 5.1 by their metric coefficients (see (5.9)). This definition does not look too motivated (apart from the fact that it is obtained simply by replacing the sine and cosine in the metric coefficients of a sphere by their hyperbolic versions). Nevertheless, there are important reasons to study hyperbolic planes. This section contains a metric introduction to two-dimensional hyperbolic geometry.

5.3.1. Motivations.

Comparison spaces. Many geometric ideas, conjectures and examples arise from studying very symmetric spaces, which are often better understood than other length spaces. Many results are formulated as "comparison theorems" stating that certain quantities in a metric space are bounded by those of a relevant comparison space. For instance, our definitions of nonpositively and nonnegatively curved spaces are based on comparisons with the Euclidean plane. The three most important types of comparison spaces are the Euclidean plane, spheres and hyperbolic planes (of course, there are other important comparison spaces, such as symmetric spaces). Apparently the reader is well acquainted with the first two types of comparison spaces: we deal with the Euclidean plane since kindergarten, and some results in the intrinsic geometry of round spheres in $\mathbb{R}^3$ are proven in high school (via the three-dimensional geometry of the ambient space $\mathbb{R}^3$). A

good knowledge of spherical geometry is a must in astronomy and geography. In contrast, geometry of the hyperbolic plane $\mathbb{H}^2$ often remains obscure and mysterious, although it by no means is more difficult than spherical geometry. This section gives an introduction to hyperbolic geometry. We will see that the spheres, the plane and the hyperbolic planes together form a nice 1-parameter family of spaces. This parameter is called *curvature*; it is zero for the Euclidean plane and $1/r^2$ for a sphere of radius r. Using this parameter, all geometric formulas can be written in such a way that they are valid for all spaces in this family. On the other hand, qualitative geometric phenomena may be drastically different for Euclidean, hyperbolic or spherical geometries.

Meta-mathematical discussion: why hyperbolic spaces. It is quite natural that Euclidean space is the first choice for a space to compare with: it may be viewed as a mathematical counterpart of the space where we derive our everyday experience from. There are several meta-mathematical reasons why we add spherical and hyperbolic spaces to the list of model spaces.

1. These are the only *isotropic* 2-dimensional spaces. A space is called isotropic if it "looks the same at every point and in every direction" (that is, for every two given unit tangent vectors, there exists an isometry whose derivative maps one of them to the other). Speculating on a physical analogy, one says that the fundamental laws of space should be the same for all locations and directions (we mean the space itself, as opposed to the influence of objects such as the Earth, etc.).

2. The idea of a straight line, which is a cornerstone notion of Euclidean geometry, has been motivated by looking at trajectories of a free motion of a particle, or by light rays. In a slightly more general situation (in the presence of a potential, or in a curved space as in general relativity) a bunch of trajectories may feature two main types of local behavior: elliptic and hyperbolic. Model examples of such behaviors are exhibited by lines in spheres and hyperbolic planes.

3. In fact, the previous remark reflects a more general phenomenon: here and there in mathematics (and physics) one sees hyperbolic and elliptic objects: hyperbolas and ellipses, hyperbolic and elliptic differential equations, sine and hyperbolic sine, and even the real component and the imaginary component of a complex number. From this viewpoint, Euclidean geometry is a borderline (parabolic) case between spherical and hyperbolic geometries. One can easily visualize a sphere and see how it looks "flatter and flatter" as its radius grows. But to forget about the family of hyperbolic spaces here is the same as to forget about complex roots of a quadratic equations. In fact, one can very well think of hyperbolic planes as spheres of imaginary radii.

4. The last but still important reason is that historically the discovery and study of hyperbolic geometry had tremendous impact on geometry and mathematical ideology in general.

Axioms and models. Random historical remarks. We are so used to the Euclidean plane that we rarely ask ourselves about the origin of the very fundamental geometric laws. There are two approaches to the definition of Euclidean geometry. Following Euclid, one says that the Euclidean plane is an object satisfying a list of axioms; the list suggested by Euclid was supposed to reflect the most basic properties of the real world given in everyday experience. There are many elementary textbooks in geometry that use this approach and systematically build Euclidean geometry out of a set of axioms (the latter are more-or-less the same Euclid's axioms formulated in accordance with the modern requirements of mathematical rigor). Such an approach is used for teaching geometry at schools in most countries, and it takes a few grades for school children to build Euclidean geometry (more or less rigorously and systematically) out of its axioms. An axiomatic approach usually seems very natural for those who have undergone such a training.

Another approach is to construct a Euclidean plane (that is, its "model") out of other mathematical objects that are believed to be better understood. For instance, one may *define* a Euclidean plane as the set of all pairs (x, y) of reals. The pairs are called points, every set of points satisfying a linear equation $ax + by + c = 0$ is pronounced to be a line, and the distance between two points is introduced by the formula $|(x_1, y_1)(x_2, y_2)| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

An axiomatic approach is convenient as it begins with listing the key properties of the object; but one may doubt why such an object exists. It is wise to keep both approaches in mind and switch between them depending on a particular consideration.

When it comes to spherical geometry, we hardly can suggest a textbook that develops it out of axioms, although a set of axioms for spherical geometry can be obtained from Euclidean axioms by a few minor modifications (and to find which axioms of Euclidean geometry should be changed to define a sphere instead of Euclidean plane is a good exercise). We usually study spheres as subsets in $\mathbb{R}^3$ on the basis of Euclidean geometry. We might axiomatically define the fundamental objects (points, lines, angles, distances) by imposing certain relations given in axioms. Instead, we traditionally begin with their models using a Euclidean round sphere: we define the points in spherical geometry to be the points of $\mathbb{R}^3$ that lie at a fixed distance from the origin; we define the lines as the great circles (central

cross-sections of the sphere). The spherical angle between two lines is defined as the Euclidean angle between the Euclidean planes containing these lines. Finally the "spherical distance" is defined to be the Euclidean angle at the origin. Note that the " spherical straight lines" do not look straight from the viewpoint of the ambient Euclidean space, and as well the spherical distance is different from the distance in the ambient space. A reader may say: "Aha, but these are just geodesics, angles and distances in the induced intrinsic metric!" This is true, and this fact had its impact on the history of hyperbolic geometry. However, we are just lucky to have such a nice model for spherical geometry—not all models can be explained in such a natural way.

After this general digression, let us turn back to hyperbolic geometry.

There is a very short axiomatic definition of the hyperbolic planes: consider Euclid's set of axioms and substitute the Fifth Postulate (stating that "*two lines parallel to the same line and having a point in common must coincide*") by its negation. Certainly, Euclid's formulations are not rigorous enough by the modern standards, so one may use their version updated by Hilbert. A space satisfying this new set of axioms is a hyperbolic plane. There is, however, a long historical path behind this definition.

The story goes back to Euclid, who demonstrated ingenious intuition by writing that after a few unsuccessful attempts to derive the fifth postulate from other axioms, he gave up since he felt it could lead him too far. Indeed, numerous attempts to prove the Fifth Postulate (based on the other postulates) led quite far! There were many "proofs", which usually drew a "contradiction", but a contradiction with the everyday intuition rather than the other axioms. For instance, one can prove that, if the Fifth Postulate were incorrect, then there is a number such that the area of every triangle is smaller than this number. Is not it a "contradiction", for we "know" that one can build as huge a triangle as she wishes? The only problem is that this "knowledge" does not follow from Euclid's axioms.

Perhaps the first mathematician who systematically studied the system of axioms with the negation of the Fifth Postulate was K. F. Gauss. (The mathematical world described by this collection of axioms is called "non-Euclidean geometry".) Gauss seriously suspected that there may be no contradiction in this axiomatic system, although many properties of non-Euclidean geometry are very different from what the common sense based on physical experience would prompt us to expect. In this respect Gauss went as far as Lobachevsky and Bolyai, and the reason why the latter two are known as "the inventors of non-Euclidean geometry" is that Gauss had not published his work. Surprisingly enough, it was just "too much knowledge" that kept Gauss from publishing his discovery: he realized that one had to

build a *model* to *prove* that there was no contradictions in non-Euclidean geometry (though neither Lobachevsky nor Bolyai had had such models, which were constructed much later by Beltrami and Poincaré). In his letters, Gauss did express his personal belief that there is no contradiction in the axioms of non-Euclidean geometry. He had an ill-fated, though an extremely wise idea of how to construct a model: he wanted to realize hyperbolic geometry as the intrinsic geometry in some surface in $\mathbb{R}^3$—the same way as spherical geometry is realized by Euclidean spheres. Gauss even found small embedded regions with desired properties (so-called pseudo-spheres), but he could not realize the whole plane. This forced him to suspect that this might be an indication that a contradiction was still hidden somewhere. It was later shown by D. Hilbert that the entire hyperbolic plane *cannot* be realized by an embedded surface. We will see that it can be represented by a Riemannian metric on the plane; the problem is that it cannot be embedded isometrically into $\mathbb{R}^3$.

Finishing this historical digression, let us mention that if one wants to develop hyperbolic geometry from its axioms, it is the same long way as we took in school to arrive at interesting results in Euclidean geometry. Perhaps having this experience once is quite enough, and one usually studies hyperbolic geometry by means of its models based on Riemannian or even Euclidean geometry. Notice, however, that if one kept track of the axioms used to derive each particular statement in Euclidean geometry, the statements can be divided into two classes: those that can be proved without involving the Fifth Postulate, and the statements that are essentially based on it. The former class of statements is called *absolute geometry*; they are true in both Euclidean and hyperbolic planes. For instance, the fact that the altitudes of a triangle intersect in one point is an absolute statement, whereas the fact that the medians in a triangle meet in one point is purely Euclidean: it does not hold in hyperbolic or spherical geometries.

5.3.2. Elementary hyperbolic geometry via Poincaré models. Let us describe the Poincaré model of hyperbolic geometry. We will postpone a proof of the fact that it is isometric to a hyperbolic plane defined before, and proceed with a relatively detailed study of the hyperbolic plane given by this model. Note that by speaking about "models" we only mean that our study of a Riemannian metric is assisted by an auxiliary Euclidean structure, and hence Riemannian objects can also be regarded from a Euclidean viewpoint.

Unlike Euclidean geometry with only *one* Euclidean plane (up to an isometry), there are *many* pairwise nonisometric hyperbolic planes (one for each negative number k, thus forming a one-parameter family). All hyperbolic planes can be obtained from each other by dilations (multiplying

a metric by a constant factor). Let us begin with the hyperbolic plane with $k = -1$, which will be referred to as the hyperbolic plane.

Poincaré model in the upper half-plane.

Definition 5.3.1. Consider the open upper half-plane
$$\mathbb{H}^2 = \{(x, y) \in \mathbb{R}^2 : y > 0\}$$
of the coordinate xy-plane equipped with a Riemannian metric given by the scalar product

$$\langle V, W \rangle_{(x,y)} = \frac{1}{y^2} \langle V, W \rangle_{\mathrm{E}}, \tag{5.12}$$

where $\langle , \rangle_{\mathrm{E}}$ is the Euclidean scalar product. Hence the metric coefficients of the hyperbolic metric are given by

$$E(x, y) = G(x, y) = \frac{1}{y^2}, \quad F = 0,$$

and the hyperbolic length structure consists of all piecewise smooth curves together with a length functional L defined by

$$L(\gamma, a, b) = \int_a^b \frac{1}{y} \left| \frac{d\gamma(t)}{dt} \right| dt, \tag{5.13}$$

where $y = y(\gamma(t))$ is just the y-coordinate of γ and $|\frac{d}{dt}\gamma(t)|$ stands for the Euclidean magnitude of the velocity vector. This length space is called the *hyperbolic* (or *Lobachevsky*) plane, and it is usually denoted by $\mathbb{H}$ or $\mathbb{H}^2$. As usual, when speaking about a length space, we mean the whole class of isometric spaces. The particular representation of this space given in this definition will be referred to as the *Poincaré model.*

Note that the hyperbolic length structure is a particular case of the conformal length structures (that is, its quadratic form $Q(V, V)$ is a scalar multiple of the Euclidean one).

Let us describe the primary objects of this geometry in terms of the Euclidean structure of the xy-plane (for elementary properties of *inversions*, see Appendix 5.3.6 at the end of this section):

- Hyperbolic lines (geodesics) are Euclidean semi-circles orthogonal to and vertical rays starting from the x-axis.
- Hyperbolic rigid motions (isometries) are translations parallel to the x-axis, symmetries w.r.t. vertical lines, dilations (homotheties) and inversions with centers in the x-axis, and their (finite) compositions. If we regard $\mathbb{H}$ as the set of complex numbers z with

$\mathrm{Im}(z) > 0$, the hyperbolic rigid motions are complex transformations of the form

$$T(z) = \frac{az+b}{cz+d},$$

where a, b, c, d are arbitrary reals with $ad - bc = 1$.

- The distance between two points $p = (0, y_1)$ and $q = (0, y_2)$ on the y-axis is

$$d(p,q) = |\ln y_1 - \ln y_2|.$$

  It will be shown below that, given two points a and b, one can choose a rigid motion that maps a and b to two points p and q in the y-axis. Then we can use the y-axis as a "ruler" since $d(a,b)$ is supposed to be equal to $d(p,q)$.
- The hyperbolic angle between two hyperbolic lines is the usual (Euclidean) angle between (the tangents to) Euclidean semi-circles (or a semi-circle and a ray) representing the lines in the Poincaré model.
- The hyperbolic area of a region Ω is defined by integration as

$$h_2(\Omega) = \int_\Omega \frac{1}{y^2}\, dxdy.$$

Certainly, a reader understands that these claims require proofs.

Warning. Dealing with the hyperbolic plane in the Poincaré model, it is often very tempting to appeal to Euclidean notions. This may be dangerous: a definition of a hyperbolic notion should be invariant under the hyperbolic isometries. For instance, Euclidean angle does persist under the hyperbolic isometries, and hence our definition of hyperbolic angle is at least not meaningless. On the other hand, Euclidean area in the Poincaré model does not have any meaning in hyperbolic geometry, for it gets distorted by hyperbolic rigid motions.

Terminological remark: To emphasize that a notion is regarded with respect to the structure of the hyperbolic plane (as opposed to the Euclidean structure of the Poincaré model), we will add an adjective "hyperbolic". For instance, we may speak about hyperbolic lines (which may be Euclidean circles), hyperbolic distances, etc.

It may seem that the hyperbolic plane defined via the Poincaré model is rather nonhomogeneous: it even has two types of lines: Euclidean semi-circles and rays. When we approach the x-axis of the Poincaré model, hyperbolic distances are huge compared to Euclidean ones; and they become small as we move towards $y \to \infty$. However, intrinsically the hyperbolic plane is quite homogeneous; this becomes obvious by looking at hyperbolic rigid motions. Indeed, given two points, it is easy to find a hyperbolic rigid

motion that maps one of the points to the other. Notice that one can do this even without using inversions! They become handy when we want to map a given ray to another given ray (to prove that hyperbolic geometry is *isotropic*).

Poincaré model in the disk. There is a convenient modification of the Poincaré model, which represents the hyperbolic plane by the interior of a disc. Apparently, the wisest strategy is to keep all models in mind and switch between them depending on the needs of a particular argument.

Passing to the model in the disc is nothing but choosing a new coordinate system in the upper half-plane. Let us apply an inversion φ with respect to the circle of a radius $\sqrt{2}$ centered at $p = (0, -1)$. We could choose any inversion whose center does not belong to $\mathbb{H}$ as well. The reason behind this particular choice is that the image of the upper half-plane is exactly the (open) disc $D = \{(x, y)\colon x^2 + y^2 = 1\}$. This model of the hyperbolic plane will be referred to as the *disc model*.

Let us see how hyperbolic objects are represented in the disc model.

- The length of a curve γ is given by

$$L(\gamma, a, b) = \int_a^b \frac{2\,|\gamma'(t)|}{1 - |\gamma(t)|^2}\, dt, \tag{5.14}$$

  where $|\gamma(t)|$ denotes the Euclidean distance from $\gamma(t)$ to the origin (the center of the disc). Hence the metric coefficients of the hyperbolic metric in the disc model in the Cartesian xy-coordinates are

$$E(x, y) = G(x, y) = \frac{4}{(1 - x^2 - y^2)^2}, \quad F(x, y) = 0.$$

- Hyperbolic lines (geodesics) are represented by the diameters of the disc and the Euclidean circular arcs orthogonal to the boundary circle of the disc.
- Hyperbolic rigid motions are the rotations about the center of the disc, the symmetries with respect to lines passing through the center of the disc, the inversions mapping the boundary circle of the disc to itself, and the products of these.
- The hyperbolic angle between two hyperbolic lines is the usual (Euclidean) angle between (the tangents to) Euclidean circular arcs (or segments) representing the lines in the disc model.
- The hyperbolic area of a region Ω is defined by integration as

$$h_2(\Omega) = \int_\Omega \frac{4}{1 - x^2 - y^2}\, dxdy.$$

Exercise 5.3.2. Describe which of the arcs and diameters representing hyperbolic lines in the disc model correspond to the vertical rays in the Poincaré model.

Exercise 5.3.3. Verify that these descriptions are correct.

Hint: To convert a hyperbolic rigid motion to the disc model, we have to conjugate it by φ. Namely, let I be a hyperbolic rigid motion given as a map in the Poincaré model. To see what it does to a point $a \in D$, we first map a back to the Poincaré model by $\varphi^{-1} = \varphi$, then apply I, and then map it back to D by φ. The formula reads $\varphi \circ I \circ \varphi$.

Exercise 5.3.4. Find a complex representation for hyperbolic rigid motions in the disc model.

Hint: Use the hint to the previous exercise together with a complex representation of the inversion I.

Exercise 5.3.5. Check that the metric coefficients of the hyperbolic metric in the disc model with respect to the standard polar coordinate system (r, ρ) in the disc are given by the following expressions:

$$E = \frac{4}{(1-r^2)^2}, \quad F = 0, \quad G = \frac{4r^2}{(1-r^2)^2}. \tag{5.15}$$

Let us introduce a new coordinate system (d, ρ), where the second coordinate of a point is the same as in polar coordinates, and the first coordinate is equal to the *hyperbolic* distance from the point to the center of the disc. This is just a direct analog of polar coordinates for the hyperbolic plane. Recall that the r-lines $(r = t, \rho)$ of the polar coordinate system are also hyperbolic lines, and hence all we have to do is to re-parameterize them by hyperbolic arc length. The hyperbolic distance from the origin to a point (r, ρ) is

$$d(r) = \int_0^r 2(1-r^2)^{-1}\, dr = -\ln\frac{1+r}{1-r}.$$

Hence the map

$$(x, y) \to (r = \frac{e^x - 1}{e^x + 1}, \ \rho = y)$$

converts the hyperbolic coordinates (d, ρ) to the polar coordinates (r, ρ).

Lemma 5.3.6. *The metric coefficients of the hyperbolic metric in the disc model with respect to* (d, ρ)*-coordinates are*

$$E = 1, \quad F = 0, \quad G = \sinh^2 d.$$

Hence (by Lemma 5.2.7) (d, ρ) *is a normal coordinate system.*

Proof. We have $E = 1$ (since the d-lines are parameterized by hyperbolic arc length), and $F = 0$ (since the ρ-lines (which are Euclidean concentric circles) and the d-lines (which are Euclidean diameters of the disc) are orthogonal (in both the Euclidean and the hyperbolic metrics, for Euclidean and hyperbolic angles are equal). Finally, we get

$$G = \left\langle \frac{\partial}{\partial \rho}, \frac{\partial}{\partial \rho} \right\rangle_{\text{hyp}} = \frac{4r^2}{(1-r^2)^2} = \sinh^2 d.$$

□

Note that the metric coefficients are given by exactly the same expression as in the formula (5.9) for $k = -1$. Hence we proved that the definition of the hyperbolic plane via the Poincaré model is equivalent to the one given in Section 5.1 by (5.9).

Hyperbolic planes of different curvatures. Together with the hyperbolic plane given by the Poincaré model, we can consider a continuum of other spaces just by multiplying the hyperbolic length by a positive constant. We will see that the Gaussian curvature of the hyperbolic plane is identically equal to -1; after rescaling its metric by multiplying it by a we will obtain a Riemannian metric of Gaussian curvature $-1/a^2$. Hence one can consider a hyperbolic plane of curvature $-k$ for each positive k.

To better understand the meaning of the procedure of rescaling a Riemannian metric, let us first apply it for the Euclidean plane. Namely let E_c be the length space whose points are points of $\mathbb{R}^2$, but the distance between every two points is multiplied by c: $d_c(x, y) = c \cdot |xy|$.

Exercise 5.3.7. Show that all E_c are isometric to each other and thus to $E_1 = \mathbb{R}^2$.

Hint: Consider the map $h_c \colon E_c \to \mathbb{R}^2$, where $h_c(x) = cx$. This map is an isometry.

This example might prompt a completely wrong idea that this procedure always leads to an isometric space. Actually, this happens only in exceptional cases. Indeed, let S_c be the length space whose points are points of the unit 2-dimensional sphere S^2, and the distance function d_c is obtained from the distance function of S^2 by multiplication by $\sqrt{c}$. Then S_c is not isometric to $S_1 = S^2$: S_c is isometric to a sphere of the radius $\sqrt{c}$. Thus we obtain the family of the spheres. The number c^{-1} is called the curvature of the sphere (and it is equal to the Gaussian curvature of its Riemannian metric; see the next chapter).

Analogously, for each positive c we consider a space $\mathbb{H}_c$ obtained from $\mathbb{H}$ by multiplying all distances by $\sqrt{c}$. The formula for its length structure

reads:

$$L(\gamma, a, b) = \sqrt{c} \int_a^b \frac{1}{y} \left| \frac{d\gamma(t)}{dt} \right| dt.$$

We will see that all spaces $\mathbb{H}_c$ are different (pairwise nonisometric). The number $-1/c^2$ is called the curvature of $\mathbb{H}_c$ (and we will also see later that it is equal to its Gaussian curvature).

Remark 5.3.8. Let us fix a number r, and let B_k be a ball of radius r in $\mathbb{H}_k$. Then the spaces B_k "converge" to B_0 as $k \to 0$, and thus on any fixed scale the hyperbolic metric with $k \to 0$ looks more and more like a Euclidean one (one can formalize this by choosing two normal coordinate systems and comparing distances between points with the same coordinates, as in the proof of Lemma 5.1.16).[1]

Remark 5.3.9. In fact, already the choice of the rigid motions of the hyperbolic plane in the Poincaré model determines the Riemannian length structure up to a multiplicative constant. Indeed, let us try to choose expressions for the metric coefficients. Since the group of isometries contains the translations along the x-axis, the expressions for metric coefficients cannot depend on the x-coordinate, and thus they are functions of y. Since Euclidean homotheties are hyperbolic isometries, the metric coefficients must be homogeneous of order -2; that is, they have to satisfy $f(cy) = f(y)/c^2$ for all positive c. Hence each metric coefficient has to be of the form const $\cdot\, y^{-2}$. Now one notices that the derivatives of the inversions that fix a given point generate all rotations around this point, and hence the hyperbolic scalar product has to be a multiple of the Euclidean one. This uniquely determines $E = G = \text{const} \cdot y^{-2}$, $F = 0$.

Rigid motions and hyperbolic lines. We want to make sure that our definitions of geometric notions introduced via the Poincaré model are consistent. First, let us verify that the hyperbolic rigid motions are isometries. Since all transformation called "the hyperbolic rigid motions" are bijective, it suffices to check that they preserve the hyperbolic length structure.

Parallel translations along the x-axis obviously leave the integrand in (5.13) unchanged, for they do not change the magnitude of γ', nor the y-coordinate. The same applies to the symmetries in vertical lines.

---

[1] It is a historical fact that Gauss tried to experimentally verify whether the physical space is non-Euclidean. He did this by measuring angles in a huge triangle (with the tops of three mountains as vertices, and using light rays as the sides of his triangle). He observed that, with available precision, the sum of the angles was π. This allowed him only to conclude that, even if that was a hyperbolic triangle, the curvature had to be very small. As a matter of fact, the measurements of Gauss did not have nearly enough precision to detect effects caused by general relativity, or otherwise he indeed would have noticed that the sum of the angles is different from π, indicating that physical space is indeed curved!

Consider a homothety with a coefficient c. Without loss of generality we assume that its center is at the origin, for we can always compose it with a translation along the x-axis. For a curve $\gamma(t) = (x(t), y(t))$, the velocity of its image $c\gamma(t) = (cx(t), cy(t))$ under the homothety is $c\gamma'(t)$. Thus the magnitude of the velocity gets multiplied by c. Since the y coordinate of $\gamma(t)$ also gets multiplied by c, the integrand in (5.13) remains unchanged.

For an inversion mapping p to q, one can see that its differential is the composition of the differential of the (unique) homothety with the same center sending p to q, and a line symmetry. Thus the above argument applies to inversions as well.

One can also see this from a straightforward computation as well. Consider an inversion I. Without loss of generality we assume that I is the inversion in the unit circle centered at the origin, for we can always achieve this by composing it with an appropriate homothety and a translation along the x-axis. The inversion I acts by the formula

$$I(x,y) = (\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2}).$$

For a curve $\gamma(t) = (x(t), y(t))$, the length of its composition with I is

$$L(I(\gamma,a,b)) = \int_a^b y(t)^{-1}\left|\frac{dI(\gamma(t))}{}dt\right| dt$$

$$= \int_a^b \left(\frac{y}{x^2+y^2}\right)^{-1} \sqrt{\left(\frac{d}{dt}\frac{x}{x^2+y^2}\right)^2 + \left(\frac{d}{dt}\frac{x}{x^2+y^2}\right)^2}\, dt = L(\gamma,a,b).$$

Exercise 5.3.10. Re-prove these statements using a complex representation

$$T(z) = \frac{az+b}{cz+d}, \qquad a,b,c,d \in \mathbb{R}, \quad ad-bc=1,$$

for a hyperbolic rigid motion T.

Exercise 5.3.11. Let us denote by $T_{a,b,c,d}$ a hyperbolic rigid motion given by a complex formula

$$T_{a,b,c,d}(z) = \frac{az+b}{cz+d}, \qquad a,b,c,d \in \mathbb{R}, \quad ad-bc=1.$$

Verify that the composition of transformations $T_{a,b,c,d}$ and $T_{a',b',c',d'}$ corresponds to the product of matrices with entries a,b,c,d and a',b',c',d':

$$T_{a,b,c,d} \circ T_{a',b',c',d'} = T_{aa'+bc',ab'+bd',ca'+dc',cb'+dd'}.$$

Hence the group of hyperbolic rigid motions is isomorphic to $SL_2(\mathbb{R})$—the group of 2×2-matrices with determinant 1.

As was mentioned already, the hyperbolic plane is homogeneous, and actually a very small part of its isometry group makes it homogeneous. Indeed, in the Poincaré model, if two points have the same y-coordinates, there exists a parallel translation along the x-axis that maps one of the points onto the other. If the y-coordinates of two points are different, then the line through these points intersects the x-axis. There is a homothety with its center at the intersection point that maps one of the points onto the other.

Definition 5.3.12. A Riemannian manifold is called *isotropic* with respect to a group of transformation G if, for every two geodesic rays, there exists a transformation from G that maps one ray to the other one.

Proposition 5.3.13. *The hyperbolic plane is isotropic (with respect to its isometry group).*

Proof. The proof becomes obvious if we use the model in the disc. Indeed, both vertices of the rays can be mapped to the center of the disc (by homogeneity). Now the rays are represented by two semi-open Euclidean segments (two radii of the disc), and there is a rotation that maps one of them to the other. □

The proposition immediately implies:

Lemma 5.3.14. *Given two points p, q in the disc model, there exists a hyperbolic rigid motion that maps p and q to two points that lie in the same diameter of the disc.*

Of course, the Euclidean plane and the spheres are isotropic as well as the hyperbolic planes, and this fundamental property is usually included in Euclid-type lists of axioms.

As a matter of fact, the spheres, the Euclidean plane, and the hyperbolic planes give an exhaustive list of two-dimensional isotropic spaces.

Exercise 5.3.15. 1. Prove the previous proposition directly in the Poincaré model (without using the model in the disc).

2. Show that the orientation-preserving isometry I that maps a given ray to a given ray is unique (as well as the one that reverses the orientation).

3. Show that our list of rigid motions of the Poincaré model is complete (that is, it includes all isometries of the hyperbolic plane).

Definition 5.3.16. A metric space (X, d) is said to be *fully homogeneous* if, for every two subsets $A, B \subset X$, and an isometry $f\colon A \to B$ (with respect to the *restrictions* of the distance function to A and B), f can be extended to an isometry of the entire X: there exists an isometry $\tilde{f}\colon X \to X$ such that $\tilde{f}|_A = f$).

Exercise 5.3.17. Give an example of a homogeneous space which is not fully homogeneous.

Exercise 5.3.18. Show that the hyperbolic plane is fully homogeneous. In particular, for two hyperbolic triangles with pairwise equal lengths of sides, there exists a rigid motion that maps one of the triangles to the other one.

Remark. Notice that, while there is no essential difference between full homogeneity and isotropy in dimension 2, full homogeneity is much stronger in higher dimensions. Can you give an example of an isotropic Riemannian manifold which is not fully homogeneous?

We will use the richness of the group of hyperbolic rigid motions to show that the hyperbolic lines are shortest paths. Notice that we will prove not only that hyperbolic lines are geodesics (and hence locally shortest paths): we will show that *every segment of a hyperbolic line* is the shortest path between its endpoints!

First of all, let us prove a particular case of this statement:

Lemma 5.3.19. *A segment that belongs to a diameter of the disc in the disc model is the unique shortest path between its points.*

Proof. The segment belongs to a d-line in the normal coordinate system (ρ, d) (see Lemma 5.3.6). Hence it is a geodesic. Now the statement follows from Lemma 5.2.9 and a trivial observation that the injectivity radius is ∞ (since the coordinates are defined in the entire disc of the disc model). □

Now, given two points p and q (in the disc model), by Lemma 5.3.14 we can map them by a rigid motion I onto two points $I(p), I(q)$ lying on the same diameter of the disc. Since I is an isometry of the hyperbolic plane, the image of a shortest path between p and q is a shortest path between $I(p)$ and $I(q)$. According to Lemma 5.3.19, the unique shortest path between $I(p)$ and $I(q)$ is the Euclidean segment $[I(p), I(q)]$. Therefore its pre-image, which is an arc of the circle passing through p and q and orthogonal to the boundary of the disc, is the unique shortest path between p and q.

Along the way we verified one of the principal axioms of both Euclidean and hyperbolic geometries: there is exactly one line passing through any pair of distinct points. Of course, this statement becomes obvious in the Poincaré model: move the two points to the same vertical line of the Poincaré model by a (hyperbolic) rigid motion. Now it is clear that two distinct points in the same vertical line cannot belong to the same circle centered at the x-axis, and hence the only line that passes through the points is the vertical ray.

Exercise 5.3.20. Draw an example of two intersecting lines that do not intersect another line (a counter-example to the Fifth Postulate).

Exercise 5.3.21. Show that, for every two intersecting lines, there exists a line that is perpendicular to one of the lines and does not intersect the other one.

Angles in Poincaré model. Let us pay attention to an important property of the Poincaré model. The identity map of the upper semi-plane $\mathbb{H}^2$ considered as a map of the Euclidean upper semi-plane onto the hyperbolic plane has a remarkable property: it preserves angles between lines; i.e., the Euclidean angle between curves is equal to the hyperbolic angle between their images. This is clear from (5.12).

The traditional term for angle-preserving maps in Riemannian geometry is *conformal maps*. Thus the identity map of the upper semi-plane onto $\mathbb{H}^2$ is a conformal map. This makes the Poincaré model very convenient: although Euclidean (straight) lines rarely remain straight with respect to the hyperbolic metric, we can still analyze angles between curves as if they were usual Euclidean angles. The disc model is also conformal, for it is obtained by applying an inversion to the Poincaré model, and inversions are conformal maps.

Remark. There are models representing the hyperbolic plane as subsets of the Euclidean plane and such that the hyperbolic lines are represented by Euclidean lines (or segments). These models, however, are never conformal, and they prove to be less convenient in most cases.

5.3.3. Ideal boundary. Recall that the boundary circle $\{(x, y):\ x^2 + y^2 = 1\}$ does not belong to the disc model. It is useful to adjoin it to the hyperbolic plane, thus obtaining a compact space homeomorphic to a disc. Let us call the boundary disc of the disc model the *ideal boundary* of the hyperbolic plane; we will denote it by Γ. The inversion that maps the disc model to the Poincaré model maps the boundary of the disc to the x-axis, and one point of the boundary disc has no image (one can imagine that it is mapped to a "point" with $y = \infty$). Hence, if working with the Poincaré model, one should think of the ideal boundary as consisting of the x-axis and one more "infinitely remote point $y = \infty$". To avoid this complication, we suggest sticking to the disc model until we give an intrinsic definition of the ideal boundary.

There is no reasonable way to extend the hyperbolic metric to the ideal boundary. However, there is a natural topology (a notion of converging sequences) in the union $\mathbb{H}^2 \cup \Gamma$, which makes it a compact space (and this compactification is finer than the one-point compactification). By definition, a sequence of points $a_i \in \mathbb{H}^2$ converges to a point $a \in \Gamma$ if it converges to a in the Euclidean topology of the disc model. Two oriented lines in the disc are said to be *asymptotic* if they are represented by circular arcs (or diameters)

whose closures have one endpoint in Γ the same. The property of lines to be asymptotic is an equivalence relation (prove this). Nonintersecting and not asymptotic lines are said to be *hyperparallels*.

Intrinsic description of the ideal boundary. A disadvantage of these definitions is that they appeal to a model, and hence their intrinsic meaning (if any) remains obscure. Here is an intrinsic way of defining Γ.

One says that two geodesics $\gamma(t)$, $\gamma_1(1)$ are asymptotic if the (hyperbolic) distance $d(\gamma(t), \gamma_1(t))$ is bounded for $t \in (0, \infty)$.

Exercise 5.3.22. 1. Prove that this definition is consistent with the one given before.

Hint: We need to estimate the hyperbolic distance between two Euclidean circular arcs meeting at the same point of the boundary circle of the disc model. Without loss of generality we can assume that this point is $(0, -1)$ (by homogeneity). Converting these lines to the Poincaré model by the inversion I, we get two vertical rays. Now this is an easy computation.

2. Prove that if geodesics $\gamma(t)$, $\gamma_1(t)$ are asymptotic, then there exists a constant t_0 such that $\lim_{t\to\infty} d(\gamma(t), \gamma_1(t + t_0)) = 0$.

3. Show that, for this choice of t_0, the limit

$$\lim_{t\to\infty} e^{-t} d(\gamma(t), \gamma_1(t + t_0))$$

is a positive (finite) number.

Hint: Use the hint to the first part.

Now we can define Γ as the set of equivalence classes of asymptotic lines. One says that a sequence of points $p_i \in H$ converges to $[\gamma] \in \Gamma$ if $d(p_i, \gamma(0)) \to \infty$ and the angle α_i between γ and the geodesic segment $[\gamma(0), p_i]$ at $\gamma(0)$ tends to 0 as $i \to \infty$.

Exercise 5.3.23. Verify that this definition is equivalent to the definition via the disc model.

For two asymptotic lines ab, ac with a common point $a \in \Gamma$, we *define* the angle between them at a to be zero. This definition is motivated by the fact that, for two points $b, c \in \mathbb{H}^2$ and a sequence $a_i \in \mathbb{H}^2$, $a_i \to a$ as $i \to \infty$, the angle $\measuredangle b a_i c$ tends to zero as $i \to \infty$ (prove this!).

Now, in addition to "bounded" triangles, we will consider *ideal triangles*. We say that a triangle is *ideal* if one or more of its vertices belong to the ideal boundary. According to our convention, the angle at such a vertex is equal to zero by definition. For example, if all vertices of a triangle are in the ideal boundary, then all its angles are equal to zero.

It is clear (from the disc model) that, for an oriented line l and a point $p \notin l$, there is exactly one line l_p^+ passing through p and asymptotic to l. If we reverse the orientation of l, we also get one line l_p^- that passes through a and is asymptotic to l with reversed orientation. All lines passing through p "between" l_p^+ and l_p^- are hyperparallel to l (i.e., do not intersect it and are not parallel to it). We suggest the reader make a sketch to visualize this observation.

Horocycles. Busemann functions.

Definition 5.3.24. Let $a = [\gamma]$ be a point in the ideal boundary Γ. A curve orthogonal to all geodesics asymptotic to γ is called a horocycle centered at a.

Exercise 5.3.25. Prove that in the Poincaré model the horocycles are represented by the circles tangent to the x-axis and the horizontal lines. Prove that in the disc model the horocycles are represented by the circles tangent to the boundary circle.

To understand the metric nature of horocycles, let us introduce a Busemann function $B_\gamma \colon \mathbb{H}^2 \to \mathbb{R}$ associated with a geodesic γ:

$$B_\gamma(p) = \lim_{t\to\infty} (d(p, \gamma(t)) - t).$$

Exercise 5.3.26. Prove that the limit does exist.

Hint: Notice that the expression $(d(p, \gamma(t)) - t)$ is monotone in t (by the triangle inequality).

For a fixed t, one can think of the function $d_t(p) = d(p, \gamma(t)) - t$ as the distance function to $\gamma(t)$ normalized by subtracting t to get zero at $\gamma(0)$. Hence a level curve $S_{\tau,t} = \{p \in H \colon d(p, \gamma(t)) - t = -\tau\}$ of d_t is a circle centered at $\gamma(t)$ and passing through a point $\gamma(\tau)$ (orthogonally to γ').

Exercise 5.3.27. Prove that the level curves $S_\tau = \{p \in H \colon B_\gamma(p) = \tau\}$ are horocycles.

Now one can think of a horocycle S_τ passing through $\gamma(\tau)$ as a limit of spheres "whose centers tend to infinity along γ". This motivates us to define a horoball as a set $D_\tau = \{p \in H \colon B_\gamma(p) \le \tau\}$.

Exercise 5.3.28. Prove that a horoball D_τ is the union of the balls $\mathrm{Ball}_{\gamma(t)}(t - \tau)$, $t \ge \tau$.

Exercise 5.3.29. What would this construction produce in the Euclidean plane? *Answer:* a semi-plane orthogonal to γ.

There is an example of a web formed by a family of asymptotic lines and a family of horocycles orthogonal to them that is very easy to visualize in the Poincaré model. It consists of the families of vertical rays and horizontal lines, which are coordinate lines in Cartesian coordinates of the xy-plane. (Note that Cartesian coordinates are *not* normal for the hyperbolic metric, for the y lines are not parameterized by the arc length).

Exercise 5.3.30. Convert Cartesian coordinates into hyperbolic normal coordinates by reparameterizing $y_1 = \varphi(y)$, and compute metric coefficients in this coordinate system.

The reader probably has already realized that a web formed by a family of asymptotic geodesics and the family of horocycles orthogonal to them can always be represented by coordinate lines in a *normal coordinate system.* In particular, horocycles form an equidistant family.

Busemann compactification. Let us describe a general construction of the Busemann compactification of a (complete locally compact) length space. This compactification is finer than the one-point compactification, and adding the ideal boundary to the hyperbolic plane is an example of this compactification.

Consider a locally compact and complete length space X. Denote by $\mathrm{Lip}(X)$ the space of all Lipschitz-1 functions on X with C^0-metric $d_{C^0}(f,g) = \sup|f-g|$ on it. Note that C^0-distance between two Lipschitz functions on a noncompact space can be infinite—but this is not a problem since our notion of a metric space allows infinite distances. If one uses a more restricted notion of a metric space, $\mathrm{Lip}(X)$ can be decomposed into a (usually a continuum) family of metric spaces with C^0-metric on each of them, and infinite supremum of the difference between any two functions from different spaces.

One can isometrically embed X into $\mathrm{Lip}(X)$ by $x \to d(x,\cdot)$. (Prove yourself that this map really is an isometry.) It is more convenient to fix a reference point $y \in X$ and define an embedding by $x \to d_x = d(x,\cdot) - d(x,y)$. This embedding puts into correspondence to a point x a distance function $d_x(z) = d(x,z) - d(x,y)$ normalized (by subtracting a constant $d(x,y)$) to make it vanish at y. Note that C^0-distance between any two distance function is finite. We say that a function $g \in \mathrm{Lip}(X)$ is a distance-like function if $d_{C^0}(g,f) < \infty$ for some (and hence for every) distance function f. An example of a distance-like function is a function $d(x,\cdot) + \mathrm{const}$.

For a noncompact X, the image of the embedding $x \to d_x = d(x,\cdot) - d(x,y)$ is not closed in $\mathrm{Lip}(X)$. Let us consider its closure, which is called the Busemann compactification of X. It contains an isometric copy of X (whose points are represented by distance functions normalized to be zero

at a reference point y), and a set of limit points of the image of X. This set of limit points, denoted by $X(\infty)$, is the *Busemann boundary* at infinity of X, and its elements are called *Busemann functions.* These are distance-like Lipschitz-one functions that arise as limits (uniform on compacts) of sequences of normalized distance functions $d_n(z) = d(z, x_n) - d(x_n, y)$.

There is another notion of the boundary at infinity (there are at least five different and useful notions of boundaries at infinity). The *ray boundary* consists of equivalence classes of rays starting from a reference point (and two rays are equivalent if they are asymptotic, that is, stay within bounded distance from each other). This set is equipped with the topology of uniform convergence on open sets. For the hyperbolic plane, ray boundary and Busemann boundary coincide (check this!).

Exercise 5.3.31. 1. What is the Busemann boundary at infinity for the Euclidean plane? For a free group with two generators (with respect to a word metric)?

2. Show that Busemann functions of rays in the hyperbolic plane are a particular case of general Busemann functions described here (for a sequence of points x_n escaping to infinity along a ray).

3. Show that, for the hyperbolic plane, its Busemann compactification coincides with its compactification we described earlier.

4. Give an example of a length space whose Busemann boundary at infinity is different from its ray boundary.

5.3.4. Gauss–Bonnet Formula. It is well-known that the sum of angles of every planar triangle is equal to π. This property is equivalent to the Fifth Postulate. (It has been known for centuries that the Fifth Postulate follows from the existence of just one triangle whose sum of angles is π.) This is a particular case of a more general *Gauss–Bonnet Formula*, which in its turn is a particular case of Theorem 6.3.17. In a certain sense, this formula gives a qualitative version of Exercise 6.5.7

As usual, by a triangle in a length space we understand three points a, b, c connected by three shortest segments $[ab]$, $[bc]$, and $[ca]$. Now we are going to consider a triangle that bounds a region T in a hyperbolic plane or a sphere (ideal triangles in a hyperbolic plane are not excluded). Then the following Gauss–Bonnet Formula holds.

Theorem 5.3.32.

$$\measuredangle a + \measuredangle b + \measuredangle c - \pi = k \cdot \mathrm{Area}(T),$$

where k is a curvature of the space.

Proof. It is enough to consider two cases: $k = 1$ and $k = -1$, and then use a dilation.

Case 1: $k = 1$. A spherical triangle can be represented as the intersection of three semi-spheres. Our argument will consist of an elementary application of the inclusion-exclusion formula.

Let a', b', c' be the points of the unit sphere opposite to a, b, c, respectively. For each pair of opposite points, say a, a', one of the two bi-gons $B(aa')$, $B'(aa')$ with the vertices a and a' covers the triangle $\triangle abc$, and the other one covers the antipodal triangle $\triangle a'b'c'$. Notice that the area of each of these bi-gons is equal to twice the angle α of $\triangle abc$ at a.

All the six bi-gons cover the entire sphere with multiplicity 1 (each point is covered by exactly one of them) except that the triangles $\triangle abc$ and $\triangle a'b'c'$ are covered three times each. By the inclusion-exclusion formula we get

$$\begin{aligned}4\pi &= \text{Area }(S^2)\\ &= 2(\text{Area }(B(aa')) + \text{Area }(B(bb')) + \text{Area }(B(cc'))) - 4\text{Area }(T).\end{aligned}$$

Together with the observation that the area of each of the bi-gons is twice the corresponding angle of $\triangle abc$ this proves case 1.

Case 2: $k = -1$. Observe that if a triangle (maybe an ideal one) is cut into two triangles and the conclusion of Theorem 5.3.32 holds for two of the three triangles, then it also holds for the third one. Every triangle can be obtained by cutting an ideal triangle off another ideal triangle. Hence it is enough to prove 5.3.32 for ideal triangles.

Consider an ideal triangle $\triangle abc$, $a \in \Gamma$. Using the Poincaré model, one can map it (by a hyperbolic rigid motion) to such an ideal triangle that two of its sides belong to two vertical rays, and two vertices lie on the circle $x^2 + y^2 = 1$ (and the third one is mapped to the "infinitely remote point $y = \infty$). An elementary (Euclidean) geometric consideration shows that if the angles of this triangle are $0 > \alpha \geq \beta > \pi$, then its vertices have Cartesian coordinates $(-\cos\alpha, \sin\alpha)$ and $(\cos\beta, \sin\beta)$.

Now the argument is based on a straightforward integration. Indeed, the hyperbolic area of the triangle T is equal to

$$\text{Area } T = \int_{-\cos\alpha}^{\cos\beta} dx \int_{\sqrt{1-x^2}}^{\infty} \frac{dy}{y^2} = \pi - \alpha - \beta.$$

□

Exercise 5.3.33. Formulate and prove an analog of the Gauss–Bonnet Formula for (hyperbolic and spherical) polygons.

Exercise 5.3.34. Show that the area of a hyperbolic triangle does not exceed π.

Remark 5.3.35. There is a remarkable trick that allows one to prove many statements in hyperbolic geometry by their spherical analogs. For example, in the proof of the Gauss–Bonnet Formula we could not apply the inclusion-exclusion formula in Case 2 since the hyperbolic plane has infinite area. However, this trick allows one to apply the result of Case 1 to claim that the Gauss–Bonnet Formula is correct in general. For instance, one can also use this trick to show that the altitudes of a hyperbolic triangle intersect in one point.

Here is the idea of the trick. There are many ways to formalize the claim that spheres, the Euclidean plane and hyperbolic planes form a one-parameter family. For instance, formulas (5.9), (5.8), and (5.7) for the metric coefficients can be represented as one formula $G(r) = (\text{Re } \sin(\sqrt{k}\, r))^2$ using complex numbers. Now if an identity that can be expressed by a formula via metric coefficients is valid for all positive values of k, then by analyticity it is true for all k (for a nonzero analytic function has only isolated zeroes). All details are left to the interested reader.

5.3.5. Selected features of hyperbolic geometry. We finish this section by formulating a few distinctive geometrical properties of the hyperbolic plane. The proofs are left to the reader. These features are mainly related to large-scale properties of the hyperbolic plane, and many of them will be generalized in Section 8.4. We deal with the hyperbolic plane (of curvature $k = 1$), leaving it as a useful exercise to figure out how our statements should be modified for other values of k.

Growth of balls. Everybody knows that the length of a Euclidean circle of radius r is $2\pi r$, and the area of the ball enclosed in this circle is πr^2. By Lemma 5.1.16, small circles and balls in a Riemannian metric look almost like their Euclidean counterparts: in particular, the length of a small hyperbolic circle is $2\pi r + o(r^2)$. In contrast, for large r hyperbolic circles and balls grow exponentially fast: the area of a circle of radius r is $\pi(e^r - e^{-r})$. Surprisingly enough, for large r the area of the ball enclosed in this circle is smaller than the length of the circle: it is $\pi(e^r + e^{-r} - 2)$. In particular, the annulus between two concentric circles of radii r and $r + 1$ contains more area than the ball enclosed in the circle of radius r.

Large triangles. First of all, the Gauss–Bonnet Formula implies that the area of a triangle is at most π. Since the area of a circle of radius 1 is greater than π, the radius of a circle inscribed in a triangle is less than 1 (of course, this is a very rough upper bound; can you give a better estimate?) This means that if one chooses three points in the hyperbolic plane (arbitrarily far apart), the three segments connecting the points "almost stick together in the center of the triangle": they intersect the same disc of radius 1! Hence

each side of a hyperbolic triangle is contained in the 1-neighborhood of the union of the two other sides. When seen from far away, a big hyperbolic triangle looks like a union of three rays emanating from the center of the triangle: the triangle is contained in the 1-neighborhood of this union. Loosely speaking, the triangle looks very "slim", as if its sides have been "sucked inside".

Using these considerations, it is easy to observe the following remarkable phenomenon. Consider a triangle Δabc, and let p be a point on the side $[bc]$. Then at least one of the triangles Δabp and Δacp is almost degenerate, in the sense that at least one of the following inequalities holds: $d(a,b)+d(b,p) \le d(a,p)+2$ or $d(a,c)+d(c,p) \le d(a,p)+2$. This property will be used to define Gromov hyperbolic (δ-hyperbolic) spaces.

Note that, unlike Euclidean geometry, there is a criterion for hyperbolic triangles to be congruent "by three angles": two triangles are congruent (one of them can be obtained from the other by a rigid motion) if each angle of one triangle is equal to the corresponding angle of the other one. On the other hand, funny enough, but there are no rectangles in hyperbolic geometry!

Morse lemma. A *quasi-geodesic* with a quasi-geodesic constant c is a curve γ such that any segment of γ is at most c times longer than the distance between its endpoints:

$$L(\gamma, t_1, t_2) \le c d(\gamma(t_1), \gamma(t_2))$$

for all t_1, t_2. Of course, c cannot be smaller than 1 unless γ is a constant curve; if $c = 1$, then γ is a shortest path. We are concerned with the following question: for a fixed quasi-geodesic constant c, how far can a quasi-geodesic deviate from a geodesic between its endpoints? In the Euclidean case this deviation can be at most const $\cdot \sqrt{d}$, where d is the distance between the endpoints of the quasi-geodesic. (Prove this and give an example when this deviation is indeed as large as const $\cdot \sqrt{d}$.)

In contrast to the Euclidean case, the Morse Lemma asserts that there exists a constant D (depending on c) such that every quasi-geodesic in the hyperbolic plane lies within distance D from the geodesic segment between its endpoints. For a proof of the Morse Lemma (in a much more general form) see Section 8.4.

There is the following striking reformulation of the Morse Lemma. Suppose that we change the Riemannian metric of the hyperbolic plane in such a way that the ratio of the two metrics remains bounded. Then every shortest path of the new metric stays within a bounded distance from a hyperbolic line. This suggests that the large-scale structure of geodesics

in the hyperbolic plane is very stable; this phenomenon is closely related to structural stability of hyperbolic (Anosov) flows.

Curvature of horocycles. A Euclidean circle of a large radius locally looks like a straight line: its curvature tends to zero as the radius grows. As opposed to the Euclidean case, (geodesic) curvature of a circle in the hyperbolic plane tends to 1 as the radius grows to infinity (see Exercise 6.4.9).

This fact makes it possible for the length of a circle to grow exponentially. Indeed, recall that the derivative of the length of a curve under an equidistant variation is equal to the integral of the geodesic curvature of the curve. The geodesic curvature of a large circle is approximately 1, and this means that the derivative of the length is approximately equal to the length itself.

Since circles passing through a fixed point at a fixed direction converge to a horocycle (as the radius tends to infinity), this means that all horocycles have geodesic curvature 1. Recall that horizontal lines in the Poincaré model are horocycles. Thus we see that intrinsically they are curved with curvature 1. Which way do they bend: up or down?

Developing this idea, let us consider a curve of constant geodesic curvature. If this curvature is greater than 1, then the curve will close up (forming a circle), exactly like a Euclidean curve of constant curvature. However, if the curvature of a curve is at most 1, the curve will not close up! Imagine that you are driving a "hyperbolic car", and you keep turning with the same curvature (less than 1). Then, against all intuition, you will nevertheless drive away, and moreover you will drive away staying within bounded distance from a straight line!

Tessellations by regular polygons. As usual, we say that a polygon is *regular* if all its sides are equal to each other and all its angles are equal to each other. In other words, an n-gon is regular if its group of isometries contains $\mathbb{Z}_n$. In the disc model the vertices of a regular Euclidean polygon centered at the origin also form the set of vertices of a regular hyperbolic polygon (certainly the sides are different, for Euclidean segments are not hyperbolic geodesics). One can see that, whereas the angles of a small hyperbolic regular n-gon are almost equal to those of a Euclidean one, in the hyperbolic case the angles tend to zero as the length of sides of the n-gon tends to infinity. In particular, if $n > 4$, one can choose a regular n-gon whose angles are $2\pi/n$. Then the entire hyperbolic plane can be tessellated by isometric copies of this polygon, with n polygons meeting at each vertex. This tessellation is similar to a tessellation of the Euclidean plane by square tiles. However, the only regular tiles that can be used to pave the Euclidean

plane are triangles, squares and hexagons, whereas the hyperbolic plane can be paved by regular n-gons for all n.

Discrete groups of hyperbolic isometries. The isometry group of the Euclidean plane does not contain many interesting discrete subgroups: every infinite discrete subgroup G contains in its turn a finite-index subgroup isomorphic to $\mathbb{Z}$ or $\mathbb{Z}^2$. If we impose an additional restriction that the quotient space $\mathbb{R}^2/G$ is a manifold, there are only two examples: $G = \mathbb{Z}^2$, in which case $\mathbb{R}^2/G$ is a torus, and an extension of $\mathbb{Z}^2$ by one line symmetry (in this case $\mathbb{R}^2/G$ is a (nonorientable!) Klein bottle). The group of hyperbolic isometries is much richer, and it contains a whole world of discrete subgroups. Recall that, by Exercise 5.3.11, the group of hyperbolic rigid motions is isomorphic to $SL_2(\mathbb{R})$. Hence for instance $SL_2(\mathbb{Z})$ acts on $\mathbb{H}$ by hyperbolic isometries.

Once we have a group G acting by isometries on a hyperbolic or Euclidean plane, there is a tessellation by polygons associated with this action. To construct this tessellation, choose a point p, and consider the set of its images $\{g(p), q \in G\}$. Then the plane is tessellated by Voronoi regions $V_{g(p)}$, where

$$V_{g(p)} = \{q\colon d(q, g(p)) \leq d(q, g'(p)) \text{ for all } g' \in G\}.$$

Tessellations of the hyperbolic plane by regular $4k$-gons, $k > 1$, can be obtained this way by groups acting by isometries without fixed points. The corresponding quotient spaces are two-dimensional closed surfaces (of higher genus) together with Riemannian metrics of constant curvature -1.

5.3.6. Appendix: inversion. By definition, an inversion $I_{p,r}$ with respect to a circle of radius r centered at $p \in \mathbb{R}^2$ maps a point q to a point $I_{p,r}(q)$ on the ray $[pq)$ uniquely determined by the relation $|pq| \cdot |pI_{p,r}(q)| = r^2$. Note that $I_{p,r}$ is not defined at p. Every inversion $I_{p,r}$ can be obtained as a composition of the inversion $I = I_{(0,0),1}$, with homothety and a parallel translation. If one regards points in $\mathbb{R}^2$ as complex numbers, the inversion I is given by a very simple formula

$$I(z) = \frac{1}{\bar{z}}, \tag{5.16}$$

where $\bar{z}$ denotes the complex conjugation. We list the properties of inversions that are important for our exposition. The proofs, which can be considered as exercises in high-school geometry (or in operations with complex numbers using (5.16)), are omitted.

1. An inversion is an involution, i.e., $I_{p,r}^2 = \text{id}$. Each inversion changes the orientation.

2. An inversion is a conformal map, i.e., it preserves angles between curves. This means that the angle between the images of two curves is equal to the angle between the curves (we need this property for circles and lines only).

3. If a circle does not pass through the center of an inversion, its image is a circle. Otherwise it is a line. If a line does not pass through the center of an inversion, its image is a circle. Otherwise it is the same line.

4. The fixed points of an inversion $I_{p,r}$ are the points of the circle of radius r centered at p (the circle of the inversion). The derivative of the inversion at each point of the circle is a line symmetry.

5.4. Sub-Riemannian Metric Structures

All length structures we considered so far were constructed by introducing new length functionals. The class of admissible paths was rather standard: we used continuous or piecewise smooth paths. This section describes a remarkable type of length structure defined by modifying a class of admissible paths. Our introduction to sub-Riemannian geometry amounts to a definition of its length structure and a brief consideration of one model example. For further reading we recommend [**BR**]. Whereas this section may be omitted for the first reading, it is worth returning to it later. We hope that some readers will be intrigued enough by it to get interested in studying sub-Riemannian geometry.

We consider the simplest case, starting from a region in $\Omega \subset \mathbb{R}^3$ with its standard Euclidean metric; replacing it by a Riemannian metric would not change anything essential but require slightly more cumbersome notation. There are, however, nontrivial phenomena that show up only in higher dimensions; they remain beyond the scope of our exposition.

Later parts of this section assume that the reader is familiar with *Lie bracket* and *differential forms.*

5.4.1. Carnot–Carathéodory metrics. In a wide class of examples admissible curves are defined by imposing restrictions on their velocity at each point. For instance, for every point p in $\Omega \in \mathbb{R}^3$, choose a two-plane $H_p \subset T_p\Omega$ (through $0 \in T_p\Omega$). Assume that H_p smoothly depends on p. This object is called a *distribution* (of two-planes), or a *two-plane field.* Let us call the planes H_p admissible. We say that a (piecewise smooth) curve γ is H-admissible if its velocity vector $\gamma'(t)$ at every point always lies in the admissible plane at the point: $\gamma'(t) \in H_{\gamma(t)}$. It may sound unbelievable, but for a generic distribution every two points in Ω can be connected by an admissible curve!

The following considerations can only add skepticism to this statement. First of all, for a two-dimensional region Ω an analog of a distribution is a line field; then every accessible curve lies in an integral curve of the line field. Hence the locus of points accessible from a given point is one-dimensional, exactly as intuition prompts.

Furthermore, let us try to look at an example when we choose admissible planes parallel to each other: for instance, we choose a horizontal plane at every point. Of course, in this case every admissible curve is confined to a horizontal plane. Hence two points with different z-coordinates cannot be connected by an admissible curve. A criterion for connectivity via admissible curves will be formulated later, see 5.4.6, but it is already clear that the planes should not be tangent to a family of surfaces (unlike vector fields, it is known that a *general* field of planes is not even locally tangent to a family of surfaces).

Now a sub-Riemannian metric structure associated with a distribution H is given by the class of H-admissible curves with Euclidean length. The resulting intrinsic metric space is often called a Carnot–Carathéodory space.

Note that Carnot–Carathéodory metrics can be regarded as limits of Riemannian metrics. Indeed, choose a quadratic form Q_p on each $T_p\Omega$ so that its kernel is H_p. Consider the following family of Riemannian metrics, which depend on a real parameter h:

$$\langle V, V\rangle_R = \langle V, V\rangle + h \cdot Q_p(V, V),$$

where $\langle V, V\rangle$ is the usual scalar square of $V \in T_p\Omega$. Denote the corresponding Riemannian distance function by $d_h(p, q)$. Loosely speaking, the term $h \cdot Q_p(V, V)$ is a penalty for going in a direction that does not belong to H_p (recall that H_p is the kernel of Q_p).

Exercise 5.4.1. Prove that $\lim_{h\to\infty} d_t(p, q)$ is the Carnot–Carathéodory distance between p and q.

The reader familiar with constraints in mechanics may notice a similarity here: to confine a particle to a surface, one introduces a potential which becomes huge as one moves away from the surface. However for a generic distribution there are no surfaces tangential to it; hence such constraints are called *nonholonomic*. A classical example of a ball rolling on a plane without skidding is briefly discussed below.

Geometric control theory. There is a slight modification of this construction. Let V_i, $i = 1, 2, \dots, k$, be several smooth vector fields on Ω. We say that a path $\gamma(t)$ is admissible if there are real-valued functions $v_i(t)$,

$i = 1, \dots, k$, such that

$$\gamma'(t) = \sum v_i(t) V_i(\gamma(t)).$$

This relation means that γ is allowed to move along any linear combination of V_i's. One can think of V_i's as available controls to direct a moving point; then v_i's would be control functions. For two vector fields V_1 and V_2 that are linearly independent at every point, one can introduce a distribution H of the planes spanned by V_1 and V_2, thus reducing this construction to the previous one. A difference arises if the vector fields happen to be linearly dependent at some points.

Exercise 5.4.2. Show that, for $V_1(x, y, z) = (1, 0, 0)$ and $V_2(x, y, z) = (0, 1, x)$, every two points are connected by an admissible path.

Geometric control: parking a car. Such classes of admissible paths arise in many applications; the fields V_i's are often called *controls*. For instance, consider a bicycle (on a plane). Riding a bicycle, a cyclist combines two ways of controlling her bicycle: steering with the handlebars and pushing pedals to move forward. On the other hand, the dimension of the space of all possible position of the bicycle is three! Indeed, a position of the rear wheel can be described by its coordinates (x, y); and we still need one number α for the angle between the bicycle and the x-axis. Thus a position of the bicycle corresponds to a point (x, y, α) in $\mathbb{R}^2 \times S^1$. This subspace is called the *configuration space* for our model. Note that, although with only two controls, the cyclist can reach every point in the 3-dimensional configuration space.

Let us describe the corresponding distribution. When a bicycle moves, its rear wheel follows the front one. In other words, the velocity of the path traced by the rear wheel remains proportional to the vector connecting the wheels. In coordinates (x, y, α) introduced above this condition is equivalent to $\dot{x} \cos \alpha = \dot{y} \sin \alpha$. In other words, the velocity vector $(\dot{x}, \dot{y}, \dot{\alpha})$ of the point moving in the configuration space at a point (x, y, α) must belong to the plane orthogonal to $(\cos \alpha, \sin \alpha, 0)$. (Here $\dot{x}$ means differentiation with respect to time t.) Geometrically, at every point of the three-dimensional configuration space $\mathbb{R}^2 \times S^1$ we have chosen a two-dimensional plane of those (tangent) vectors which correspond to the velocities of actually possible motions of the bicycle: the other vectors represent dragging the bicycle in its side direction.

To produce a length structure, we need to fix a length function, though its particular choice is not so important here: the restriction of the class of admissible curves results in very unusual qualitative properties of the induced sub-Riemannian metric, whereas modifications of the length function cause relatively modest changes of the length structure. For instance, let us

use a Euclidean length given by

$$L(\gamma, a, b) = \int \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{\alpha}^2}\, dt.$$

Let's try to informally discuss the sub-Riemannian metric arising in this example. Consider two points which are very close to each other. If the vector between the two close points "lies close to an admissible direction", one should not expect that the new distance is essentially different from Euclidean one. But if the vector between the points is perpendicular to admissible planes, then an admissible path between the points has to go a long way; for such points the distance is stretched. Everybody observes this when doing parallel parking (and our model can be applied to cars as well as to bicycles): you want to displace your car just one foot to its side direction, but you have to drive 50 feet moving it back and forth. Perhaps parking a car is not such a difficult task, but parking a car with two trailers requires very advanced mastery. Geometrically, a car with two trailers gives rise to the same type of intrinsic metric, just the dimension of the configuration space is higher, while there are still only two controls.

Nonholonomic mechanical constraint: rolling a ball. Another example of the same kind is rolling a ball. A position of a ball lying on a plane requires five coordinates: two reals to characterize the point in the plane, another two coordinates to characterize the point of the ball which touches the plane, and the last one for spinning the ball around its vertical axis. When one rolls the ball without sliding, there are only three admissible directions (controls): two to choose a direction where to roll it and the third one for spinning it. Still, one can get to any position regardless of the initial position: the reader can check that experimentally rolling a globe on the floor. Nevertheless, for certain positions which are visually close, it takes quite a bit of rolling to get from one to the other: position the globe such that it touches the floor by Washington, and try to roll it back to the same point on the floor but such that now its bottom point is at New York.

5.4.2. Connectibility and Ball-Box Theorem.

Lie bracket. Here we briefly recall the notion of Lie bracket; one can find details and proofs in many books (see for example [**BC**]).

Two main objects associated with a smooth vector field V are the corresponding derivation of smooth functions $f \to Vf$ and the one-parameter group of diffeomorphisms φ_V^t generated by V. Recall that, for each t, φ_V^t is a diffeomorphism, and for a fixed p,

$$\frac{d}{dt}\varphi_V^t(p) = V(\varphi_V^t(p)).$$

In other words, for each x, $\gamma(t) = \varphi_V^t(p)$ is an integral curve of V that passes through p at $t = 0$. (A curve γ is an *integral curve* of a vector field V if $\frac{d}{dt}\gamma(t) = V(\gamma(t))$.)

Here are three equivalent definitions of Lie bracket $[V, W]$ of vector fields V, W.

Definition 5.4.3. Lie bracket $[V, W]$ is a vector field defined by the condition: $[V, W]f = VWf - WVf$ for every smooth function f.

Of course, one has to prove that there exists a unique vector field $[V, W]$ that differentiates functions according to the formula in the definition. The definition immediately implies that Lie bracket is linear in each argument and skew-symmetric. It is easy to check that the following property holds: $[fV, W] = f[V, W] - (Wf)V$, where f is a smooth function.

The next definition is especially important in our geometric context:

Definition 5.4.4. $[V, W](p) = \frac{d}{dt}\Big|_{t=0+}(\varphi_V^{\sqrt{t}} \circ \varphi_W^{\sqrt{t}} \circ \varphi_V^{-\sqrt{t}} \circ \varphi_W^{-\sqrt{t}}(p))$.

To better understand its geometric meaning, one can use a coordinate chart (to be able to "subtract points") and rewrite it in a less invariant form as

$$[V, W](p) = \lim_{t\to 0} \frac{1}{t^2}(\varphi_V^t \circ \varphi_W^t \circ \varphi_V^{-t} \circ \varphi_W^{-t}(p) - p).$$

We suggest the reader make a sketch with a "quadrilateral" formed by four segments of integral curves of V and W traversed back and forth. The "quadrilateral" has its first vertex at p, then its first side follows the integral curve of W backwards for time t, and so on. This "quadrilateral" does not quite "close up" (unless we are dealing with coordinate vector fields, in which case we actually get a closed "quadrilateral"): when one follows all four sides of the "quadrilateral", she gets to a point p' close to p; this displacement $p' - p$ is of second order in t; dividing it by t^2 and sending $t \to 0$ gives the Lie bracket.

The last definition stresses similarity between Lie brackets and usual derivatives:

Definition 5.4.5.

$$[V, W](p) = \lim_{t\to 0} \frac{1}{t}(d\varphi_V^{-t} W(\varphi_V^t(p)) - W(p)) = \frac{d}{dt}\Big|_{t=0} d\varphi_V^{-t} W(\gamma(t)),$$

where γ is the integral curve of v with $\gamma(0) = p$.

The expression on the right-hand side of the formula in this definition is called the *Lie derivative* of W with respect to the flow φ_V^t. To justify this name, imagine that you follow the integral curve of V passing through p and study how W changes: you want to compare $W(p) = W(\gamma(0))$ and

$W(\gamma(t))$. These vectors belong to different tangent spaces, so to subtract them we map $W(\gamma(t))$ to the tangent space at p by the differential $d\varphi_V^{-t}$ of φ_V^{-t}. Then, as usual, we subtract $W(p)$ from it, divide the difference by t and send t to 0.

Chow–Rashevsky's Theorem. In this section we give a sufficient condition a Carnot–Carathéodory metric to be finite. This criterion is known as Chow–Rashevsky Connectivity Theorem. Here is its simplest 3-dimensional case:

Theorem 5.4.6. *Let $\Omega \subset \mathbb{R}^3$ be a connected region, and V_1, V_2 two smooth vector fields such that $V_1(p)$, $V_2(p)$, and their Lie bracket $[V_1, V_2](p)$ form a basis of $T_p\Omega$ for every $p \in \Omega$. Then every two points $p, q \in \Omega$ can be connected by an admissible path, that is, a piecewise smooth path γ whose velocity vector at every (smooth) point is a linear combination of V_1 and V_2: $\gamma'(t) = v_1(t)V_1(\gamma(t)) + v_2(t)V_2(\gamma(t))$.*

Proof. Let us prove that, for every $p \in \Omega$, the set of points that can be connected with p by admissible paths is open.

Let φ_1^t, φ_2^t be the one-parameter groups of diffeomorphisms generated by V_1, V_2: $\frac{d}{dt}\varphi_i^t(p) = V_i(\varphi_i^t(p))$, $i = 1, 2$. Recall that

$$[V_1, V_2](p) = \frac{d}{dt}\bigg|_{t=0+} (\varphi_1^{\sqrt{t}} \circ \varphi_2^{\sqrt{t}} \circ \varphi_1^{-\sqrt{t}} \circ \varphi_2^{-\sqrt{t}}(p)).$$

Consider a map $F\colon \mathbb{R}^3 \to \Omega$:

$$F(u, v, \tau) = \varphi_1^u \circ \varphi_2^v \circ \varphi_1^{\sqrt{|\tau|}} \circ \varphi_2^{sign(\tau)\sqrt{|\tau|}} \circ \varphi_1^{-\sqrt{|\tau|}} \circ \varphi_2^{-sign(\tau)\sqrt{|\tau|}}(p). \tag{5.17}$$

This map is at least C^1-smooth (check this!) and

$$\tfrac{\partial}{\partial u}F(p) = V_1(p), \quad \tfrac{\partial}{\partial v}F(p) = V_2(p), \quad \tfrac{\partial}{\partial \tau}F(p) = [V_1, V_2](p).$$

Hence by the Inverse Function Theorem, the image of F contains a neighborhood of p. Now it remains to observe that a point $F(u, v, \tau)$ is connected with p by the following admissible path $\gamma\colon [0, 6] \to \Omega$, which alternates moving along V_1 and V_2 (that is, it is built out of segments of integral curves of V_1 and V_2):

$$\begin{aligned}
\gamma(t) &= \varphi_2^{-\sqrt{\tau}t}(p), \quad t \in [0, 1],\\
\gamma(t) &= \varphi_1^{-\sqrt{\tau}(t-1)} \circ \varphi_2^{-\sqrt{\tau}}(p), \quad t \in [1, 2],\\
\gamma(t) &= \varphi_2^{\sqrt{\tau}(t-2)} \circ \varphi_1^{-\sqrt{\tau}} \circ \varphi_2^{-\sqrt{\tau}}(p), \quad t \in [2, 3],\\
\gamma(t) &= \varphi_1^{\sqrt{\tau}(t-3)} \circ \varphi_2^{\sqrt{\tau}} \circ \varphi_1^{-\sqrt{\tau}} \circ \varphi_2^{-\sqrt{\tau}}(p), \quad t \in [3, 4],\\
\gamma(t) &= \varphi_2^{v(t-4)} \circ \varphi_1^{\sqrt{\tau}} \circ \varphi_2^{\sqrt{\tau}} \circ \varphi_1^{-\sqrt{\tau}} \circ \varphi_2^{-\sqrt{\tau}}(p), \quad t \in [4, 5],\\
\gamma(t) &= \varphi^{u(t-5)} \circ \varphi_2^{v} \circ \varphi_1^{\sqrt{\tau}} \circ \varphi_2^{\sqrt{\tau}} \circ \varphi_1^{-\sqrt{\tau}} \circ \varphi_2^{-\sqrt{\tau}}(p), t \in [5, 6].
\end{aligned}$$

Hence every point in the image of F can be connected with p; since the image of F contains a neighborhood of p, its connectivity component is open.

Since connectivity components are open and either coincide or do not intersect, the theorem immediately follows from connectedness of Ω. □

Exercise 5.4.7. Show that under the assumptions of Theorem 5.4.6 every two points in Ω can be connected by a *smooth* admissible path (as opposed to a piecewise smooth one).

In full generality the Chow–Rashevsky Theorem reads as follows: let V_i, $i = 1, 2, \ldots, k$, be smooth vector fields on a connected manifold Ω. Assume that the V_i's, their Lie brackets, and their iterated Lie brackets (such as $[[[V_1, V_2], V_3], V_4])$ span $T_p\Omega$ for every p. Then every two points can be connected by an admissible curve.

Chow's Theorem for distributions can be formulated in the following elegant way: Let $H_p \in T_p\Omega$ be a smooth distribution of linear subspaces on a connected manifold Ω. Denote by $\mathfrak{H}$ the space of vector fields contained in the distribution, that is, $V \in \mathfrak{H}$ if $V(p) \in H_p$ for all p. Assume that the Lie sub-algebra of vector fields on Ω generated by $\mathfrak{H}$ is the whole algebra of vector fields. Then every two points in Ω can be connected by an admissible path. Note that if iterated Lie brackets of vector fields from $\mathfrak{H}$ generate all vector fields, then it is enough to use at most $n = \dim(\Omega)$ iterations.

Exercise 5.4.8. In each of the following examples check that the assumptions of Chow's Theorem are satisfied, and hence the corresponding Carnot–Carathéodory metrics are finite.

(1) $V_1 = (1, 0), V_2 = (y, x)$ on $\mathbb{R}^2$ in Cartesian coordinates (x, y). Can you sketch a small sub-Riemannian ball centered at $(0, 0)$?

(2) $V_1 = (1, 0, 0), V_2 = (0, 1, x)$ on $\mathbb{R}^3$ in Cartesian coordinates (x, y, z),

(3) $V_1 = (0, 0, 1), V_2 = (\sin\alpha,\ -\cos\alpha,\ 0)$ on a three-dimensional cylinder $\mathbb{R}^2 \times S^1$ with coordinates $(x,\ y,\ \alpha)$ (the distribution of the admissible motions of a bicycle in the example above),

(4) V_1,V_2 are any two linearly-independent left invariant fields on $SO(3)$.

Exercise 5.4.9. Prove the general Chow–Rashevsky Theorem.

Ball-Box Theorem. Now we want to get some insight into infinitesimal behavior of a sub-Riemannian metric generated by a distribution H_p in Ω. We will restrict ourselves to the case $\dim\Omega = 3$; unlike Chow's Theorem, whose general version does not involve new ideas and requires only more tedious and cumbersome notations, even the formulation of the Ball-Box Theorem in higher dimensions relies on a more delicate technique (of privileged coordinates). However, already the three-dimensional Ball-Box Theorem gives a pretty good geometric idea of what is going on.

Denote by P_r^c the parallelepiped ("box")

$$[-cr, cr] \times [-cr, cr] \times [-cr^2, cr^2] \subset \mathbb{R}^3.$$

Theorem 5.4.10. *Let V_1, V_2 be two smooth vector fields in a neighborhood of the origin in $\mathbb{R}^3$, and let H_p be the corresponding distribution of two-dimensional planes spanned by V_1, V_2; assume that $H_{(0,0,0)}$ is the xy-plane $z = 0$. Suppose that the assumptions of Chow's Theorem are satisfied. Let d be the sub-Riemannian metric induced by the distribution H. Denote by B_r the d-ball of radius r centered at the origin. Then there are constants $0 < c < C$ such that, for all sufficiently small r, one has*

$$P_r^c \subset B_r \subset P_r^C.$$

Hence in the xy-directions B_r looks approximately like a Euclidean ball, whereas in the z-direction it is very thin (the length of its intersection with the z-axis is of order r^2). One can think of this ball as approximately a Euclidean ball squeezed in r times in the z-direction. Note, however, that it is not known even whether it is a topological ball for all sufficiently small r!

Proof. I. First we prove that there is a positive c such that $P_r^c \subset B_r$ for all sufficiently small r. To prove this inclusion, it suffices to show that every point from P_r^c can be connected with the origin by an admissible path of length at most r. Notice that $W = [V_1, V_2](0,0,0)$ does *not* lie in the xy-plane (by Chow's condition).

We will use the map F (for $p = (0,0,0)$) defined in (5.17) and an admissible curve γ connecting $(0,0,0)$ and $F(u,v,\tau)$ constructed in the proof of the Chow–Rashevsky Theorem 5.4.6. Let us estimate the length of γ. Let M be such that $|V_1(p)| \leq M$, $|V_2(p)| \leq M$ for all p in a neighborhood of the origin. Then the length of γ is bounded by $M|u| + M|v| + 4M\sqrt{|\tau|}$. Indeed, the velocity of γ for the intervals $t \in [0,1]$, $t \in [1,2]$, $t \in [2,3]$, $t \in [3,4]$ is $-\sqrt{|\tau|}V_2$, $-\sqrt{|\tau|}V_1$, $\sqrt{|\tau|}V_1$, $\sqrt{|\tau|}V_2$ respectively. Hence its speed for $t \leq 4$ is bounded by $\sqrt{|\tau|}M$, and thus the length of $\gamma_{|[0,4]}$ is at most $4M\sqrt{|\tau|}$. Analogously for $t \in [4,5]$ and $t \in [5,6]$, the velocity of γ is vV_1 and uV_2 respectively, and hence the length of these segments of γ is at most $M|u| + M|v|$. Thus we conclude that the Carnot–Carathéodory distance from the origin to $F(u,v,\tau)$ is at most $M(|u| + |v| + 4\sqrt{|\tau|})$. Thus it remains to show that (there exists a positive c such that) P_r^c is contained in the image under F of the set $\{(u,v,\tau)\colon |u| + |v| + 4\sqrt{|\tau|} \leq r/M\}$. This image in its turn contains the image of the rectangle

$$\{(u,v,\tau)\colon |u| \leq \delta r, |v| \leq \delta r, |\tau| \leq \delta^2 r^2\},$$

where $\delta = 1/6M$.

Now all is left is to recall that

$$\frac{\partial}{\partial u}F(0,0,0) = V_1(0,0,0),\ \frac{\partial}{\partial v}F(0,0,0) = V_2(0,0,0),\ \frac{\partial}{\partial \tau}F(p) = W.$$

Since $V_1(0,0,0)$, $V_2(0,0,0)$, W are linearly independent, there is a constant $c > 0$ such that

$$P_r^c \subset F\left([-\delta r, \delta r] \times [-\delta r, \delta r] \times [-\delta^2 r^2, \delta^2 r^2]\right),$$

provided that r is small enough. This proves part I.

II. Now we need to show that there is a positive C such that $B_r \subset P_r^C$ for all sufficiently small r; in other words, we need to estimate Carnot–Carathéodory distance from below. Let γ be an admissible path connecting the origin and (u, v, τ). We want to estimate from below the (Euclidean) length L of γ. It is enough to show that

$$CL \geq \max(|u|,\ |v|,\ \sqrt{|\tau|}). \tag{5.18}$$

Let us choose (and fix) a nonvanishing differential 1-form ω in a neighborhood of the origin such that $\omega = dz$ at the origin, and $\omega(V_1) = \omega(V_2) = 0$; i.e., the kernel of ω at every point p coincides with $H_p = \mathrm{span}(V_1(p), V_2(p))$. Let $\omega = dz + \nu$. Then $|\nu_p| \leq A|p|$ for some A since ν vanishes at the origin. Choose M such that $||d\omega|| \leq M$ in a neighborhood of the origin.

Note that L is obviously greater than the Euclidean distance from the origin to (u, v, τ). Hence $2L \geq (|u| + |v|)$. Thus, for C, say, greater than $100\sqrt{1+A}$, the inequality (5.18) is automatically satisfied for points (u, v, τ) with $\sqrt{\tau|} \geq 2\sqrt{1+A}\,(|u| + |v|)$. We will choose $C > 100(1 + A)$, and hence it is enough to handle the points with $\sqrt{|\tau|} \geq 2\sqrt{1+A}\,(|u| + |v|)$, that is, with $|\tau| \geq 4(1 + A)(|u| + |v|)^2 \geq 2(1 + A)(u^2 + v^2)$.

Form a cycle α by completing γ to a closed curve by connecting (u, v, τ) with the origin by a segment σ. Then

$$\int_\alpha \omega = \int_\gamma \omega + \int_\sigma \omega.$$

Notice that $\int_\gamma \omega = 0$ since γ' is a linear combination of V_1 and V_2, and therefore $\omega(\gamma') = 0$. Thus

$$\left|\int_\alpha \omega\right| = \left|\int_\sigma \omega\right| = \left|\int_\sigma (dz + \nu)\right| = |-\tau + \int_\sigma \nu| \geq |\tau| - A(u^2 + v^2 + \tau^2).$$

Recalling that we need to deal only with points with $|\tau| \geq 2(1+A)(u^2+v^2)$, we see that for such points

$$\left|\int_\alpha \omega\right| \geq \frac{1}{2}|\tau| - A\tau^2 \geq \frac{1}{3}|\tau| \tag{5.19}$$

for all sufficiently small τ.

Now we will make use of an isoperimetric inequality in $\mathbb{R}^3$; indeed, the length of the cycle α is at most $2L$, and hence it can be represented as the boundary of a two-chain β whose area is at most $(1/\pi)L^2$. ($1/\pi$ is a sharp constant, and the proof of this statement may be tricky; for our argument a constant, say, 10, would do as well, and the corresponding crude isoperimetric inequality is a simple exercise.) Thus choose a β with $\partial\beta = \alpha$ and $||\beta|| \le 10L^2$. By Stokes' formula,

$$|\int_\alpha \omega| = |\int_\beta d\omega| \le M||\beta|| \le 10ML^2.$$

Together with (5.19), this yields $L^2 \ge |\tau|/(30M)$. Recall that we are dealing with points with $\sqrt{|\tau|} \ge 2\sqrt{1+A}\,(|u|+|v|)$, and in particular $\sqrt{|\tau|} \ge \max(|u|,|v|)$. Thus we get that

$$\sqrt{30M}\,L \ge \sqrt{|\tau|} = \max(|u|,\,|v|,\,\sqrt{|\tau|}),$$

which proves the theorem (for $C = \max(\sqrt{30M},\,100(1+A))$). □

Exercise 5.4.11. Find the Hausdorff dimension of a (finite) Carnot–Carathéodory metric induced by a two-dimensional distribution in $\mathbb{R}^3$.

Answer: 4.

5.4.3. Model example: Connections in fiber bundles.

Model example. We complete our discussion of sub-Riemannian length structures by looking at a model example. This example will allow us to see many important features of Carnot–Carathéodory spaces, as well as to demonstrate important connections between sub-Riemannian geometry and other concepts.

The example is given by the following distribution H_p of 2-planes in $\mathbb{R}^3$: $H_{(x,y,x)} = \operatorname{span}(V(x,y,z), W(x,y,z))$, where $V(x,y,z) = (1,0,0)$, $W(x,y,z) = (0,1,x)$. This distribution can be regarded as a linearized approximation to the one describing the admissible motions of a bicycle. The Lie bracket of V and W is $[V,W] \equiv (0,0,1)$. Indeed,

$$Vf = \frac{\partial f}{\partial x}, \quad Wf = \frac{\partial f}{\partial y} + x\frac{\partial f}{\partial z}, \quad VWf - VWf = \frac{\partial f}{\partial z} = (0,0,1)f.$$

Hence the conditions of the Chow–Rashevsky Theorem are satisfied: every two points can be connected by an admissible curve. For instance, any choice of a finite length structure on smooth curves in $\mathbb{R}^3$, H_p induces a finite sub-Riemannian length structure. We could use the usual Euclidean length of curves, but it is more convenient to define the length of an admissible curve $\gamma(t) = (x(t), y(t), z(t))$ by

$$L(\gamma) = \int \sqrt{x'^2(t) + y'^{2(t)}}\,dt. \tag{5.20}$$

Note that admissible curves cannot go vertically, and hence the integrand in this formula never vanishes.

We will look at this example in various clothing, which will give different insights into the geometry of sub-Riemannian length structures.

Connections in fiber bundles. Let us think of $\mathbb{R}^3$ as $B \times F$, where the base $B = \mathbb{R}^2$ is the xy-plane and the fiber $F = \mathbb{R}$ is the z-axis. Denote the natural projection to the B-factor (base) by π. Note that we have chosen our length structure in such a way that this projection preserves the length of admissible curves; this is a reason why we preferred the length structure given by (5.20) to the Euclidean one.

The results of this section apply (after proper modifications) to a general distribution on a fiber bundle; we use the existence of a group of translations (along the z-axis) that preserve the distribution, and the fact that (the derivative of) the projection restricted to each plane of the distribution is bijective. Restricting ourselves to one-dimensional fibers allows us to avoid vector-valued forms, but apart from that more general case is not essentially more difficult.

The most important feature of this situation is that every smooth curve in the base can be lifted to $\mathbb{R}^3$ as an admissible path. Once the lift of one point is fixed, the lifting of the whole curve is unique. Indeed, the restriction of the differential $d_p\pi$ of the projection to the base to H_p is a linear isomorphism; hence for a vector field X tangent to the base B, there is a unique lift of X, that is, a vector field $\tilde{X}$ that lies in the distribution H and $d\pi(\tilde{X}) = X$. From here, it is easy to see that given a smooth path $\gamma\colon \mathbb{R} \to B$, and a point p with $\pi(p) = \gamma(0)$, there exists a unique *lift* of γ, that is, an admissible curve $\tilde{\gamma}$ with $\tilde{\gamma}(0) = p$ and $\pi(\tilde{\gamma}) = \gamma$. Of course, all admissible curves come this way. Thus if one wants to connect two points $p, q \in \mathbb{R}^3$ by an admissible curves, one can present it by drawing its projection connecting the projections p_0, q_0 of the points to the base. Notice, however, that if one connects p_0 and q_0 by some path $\gamma\colon [0,1] \to B$, its lift $\tilde{\gamma}$ with $\tilde{\gamma}(0) = p$ will connect p with some point in the fiber containing q, $\tilde{\gamma}(1) \in q_0 \times F$, though there is no reason to expect that it will be just q. This suggests the following simple but useful definition.

Holonomy group. For a path $\gamma\colon [0,1] \to B$ with $\pi(\gamma(0)) = p$, $\pi(\gamma(1)) = q$ define by $G_\gamma\colon F \times \{p\} \to F \times \{q\}$ the following transformation: $G_\gamma(a) = b$ if $\tilde{\gamma}(1) = b$ for the lift $\tilde{\gamma}$ of γ with $\tilde{\gamma}(0) = a$.

Exercise 5.4.12. Show that each G_γ is a diffeomorphism.

Exercise 5.4.13. Show that, for a fixed $p \in B$, the collection of G_γ for the loops γ with $\gamma(0) = \gamma(1) = p$ forms a group of diffeomorphisms of

$F = F \times \{p\}$, and concatenation of paths corresponds to composition of transformations. This group is called the holonomy group at p.

Note that $\mathbb{R}$ acts on $\mathbb{R}^3$ by translations in the z-direction: $g_a(x, y, z) = (x, y, z + a)$. Both the distribution H and a lift $\tilde{X}$ are invariant under this action.

Exercise 5.4.14. Verify that for our distribution every G_γ is just a translation.

Holonomy groups at different points p and q are isomorphic, and every path γ between p and q gives rise to an isomorphism: for a loop σ at p, we see that $G_\sigma \to G_\gamma^{-1} \circ G_\sigma \circ G_\gamma$ comes from the loop formed by traversing γ, then σ and then again γ in the reversed direction. This construction actually establishes a conjugacy between actions of holonomy groups at different points.

Now the problem of finding an admissible path connecting two points p_1 and q_1 reduces to finding a path γ connecting their projections $p = \pi(p_1)$ and $q = \pi(q_1)$ with $G_\gamma(p_1) = q_1$. In particular, every two points can be connected if the holonomy group at a point (and hence at every point) acts transitively (verify this assertion!)

Curvature form. The following construction plays a crucial role in understanding holonomy groups.

Given a vector $X \in T_{(x,y,z)}\mathbb{R}^3$, it can be uniquely represented by $X_H + X_V$, where $X_H \in H_{(x,y,x)}$, and $\mathrm{pr}(H_V) = 0$. The vectors X_H and H_V can be called the horizontal and the vertical components of X; X_V is tangential to the copy of the fiber F (a vertical line) passing through (x, y, z). Of course, a curve is admissible if the vertical component of its velocity is identically zero.

Now let us consider two vector fields X, Y on the base, and lift them to $\tilde{X}$, $\tilde{Y}$. Denote by $\omega(X, Y)$ the vertical component of the Lie bracket $[\tilde{X}, \tilde{Y}]$. $\omega(X, Y)$ is a vector field tangent to the fibers; since in our case all fibers are identified with $\mathbb{R}$, $\omega(X, Y)$ can be regarded as just a real-valued function.

Exercise 5.4.15. Prove that $\omega(X, Y)(p)$ depends only on *vectors* $X(\pi(p))$ and $Y(\pi(p))$. Show that ω is a two-form on the base.

Hint: Note that the vertical component of X is nothing but $\langle X, Z\rangle$, where $Z = (0, 0, 1)$.

This may be somewhat surprising, for the Lie bracket used to define Z involves derivatives; this is similar to the situation with the curvature tensor (see Chapter 6, 6.3.3), which is defined using vector fields, but at the end it

happens to depend only on their values at one point. As a matter of fact, the curvature tensor is a particular case of this construction.

The geometric meaning of the curvature form ω is given by the following important exercise:

Exercise 5.4.16. Let γ be a loop at p: $\gamma(0) = \gamma(1) = p$. Assume that γ bounds a simple region Γ, whose orientation agrees with that of γ. Then the transformation G_γ is a translation $z \to z + A$, where

$$A = \int_\Gamma \omega.$$

Hint: There is a straightforward hand-crafted argument, which is very good to reveal the geometric meaning of the construction. One considers two commuting vector fields X and Y, and first proves the statement for an infinitesimal rectangle formed by integral curves of the fields just from the definition of the Lie bracket. Then one replace the projection of γ by a piecewise path consisting of integral curves of X or Y. Another proof, which is much shorter but not that visual, can be conducted analogously to the reasoning in the section "Contact Viewpoint" below.

Shortest path and isoperimetric problem. Finally, let us look at our particular distribution $V(x, y, z) = (1, 0, 0)$, $W(x, y, z) = (0, 1, x)$. To apply ω to $X = (1, 0)$, $Y = (0, 1)$, we note that their lifts are just V and W; there Lie bracket $[V, W] = (0, 0, 1)$ has been computed at the beginning of this section; hence $\omega(X, Y) = 1$, and therefore ω is just the standard area form. This allows us to explicitly find the shortest paths of this sub-Riemannian structure. For instance, let us connect $p = (0, 0, 0)$ and $q = (0, 0, z)$ by a shortest path. The projection of any curve connecting these points is a loop at $(0, 0)$ enclosing oriented area z, and the length of this projection is the same as the length of the curve (by (5.20)). Hence the shortest path between the points is a lift of the shortest closed curve enclosing area z. This is an isoperimetric problem, and it is known that the shortest curve enclosing a given area is a circle. Hence shortest paths connecting p and q are lifts of circles of a radius $\pi^{-1}\sqrt{z}$ passing through the origin. Note that, unlike any Riemannian metric, there are infinitely many different shortest paths between p and q for every value of z.

Exercise 5.4.17. Show that every shortest path in this sub-Riemannian structure is a lift of a circular arc. Note that shortest paths may split: there are infinitely many shortest paths emanating from every point in a given (admissible) direction.

Exercise 5.4.18. Using the description of shortest paths, give an explicit formula for the sub-Riemannian distance function, and explicitly verify the conclusion of the Ball-Box Theorem.

Contact viewpoint. The distribution H_p can be regarded as a contact structure: it is the distribution of kernels of a contact 1-form $\omega_1 = xdy - dz$. Let us use an argument analogous to the one exploited to obtain a lower bound in the Ball-Box Theorem. As a model example, let us again consider an admissible path γ connecting two points $p = (0,0,0)$ and $q = (0,0,z)$. We can form a 1-cycle α by closing γ by a segment qp. Let β be a (smooth) two-chain whose boundary is α. By Stokes' Formula,

$$\int_\alpha \omega_1 = \int_\beta d\omega_1.$$

The exterior differential of ω_1 is $d\omega_1 = \omega_2 = dx \wedge dy$, which is nothing but the oriented area of the projection to the xy-plane. Hence $\int_\alpha \omega_1$ measures the oriented area enclosed in the projection of γ (for the segment qp projects to a point). On the other hand,

$$\int_\alpha \omega_1 = \int_\gamma \omega_1 + \int_{qp} \omega_1 = 0 + z.$$

Hence we obtain the same conclusion as at the end of the previous section: the shortest curve connecting p and q projects to the shortest curve bounding a region of area z.

Heisenberg group. Let us introduce the following group structure on $\mathbb{R}^3$:

$$(x, y, z) \cdot (x_1, y_1, z_1) = (x + x_1, y + y_1, z + z_1 + xy_1).$$

This is "almost" the usual addition, with a "twisting term" xy_1 in the z-component of the result. Note H is a left-invariant distribution with respect to this group structure. Indeed, a left translation $L_{(x_0,y_0,z_0)}$ that sends $(0,0,0)$ to a point (x_0, y_0, z_0) is a multiplication by (x_0, y_0, z_0):

$$L_{(x_0,y_0,z_0)}(x, y, z) = (x + x_0, y + y_0, z + z_0 + x_0 y),$$

and its differential $dL_{(x_0,y_0,z_0)}$ acts by $L_{(x_0,y_0,z_0)}(x, y, z) = (x, y, x_0 y)$. Thus

$$dL_{(x_0,y_0,z_0)}(1,0,0) = (1,0,0) = V(x_0, y_0, z_0),$$

$$dL_{(x_0,y_0,z_0)}(0,1,0) = (0,1,x_0) = W(x_0, y_0, z_0).$$

Moreover, the length structure given by (5.20) is also left-invariant (verify this!) Thus H_p determines a left-invariant Carnot–Carathéodory metric d_H on the Heisenberg group.

It is interesting to compare this metric with a left-invariant Riemannian metric d whose quadratic form at the origin $(0,0,0)$ coincides with the standard coordinate Euclidean structure.

Exercise 5.4.19. Show that d and d_H are asymptotically the same:

$$\lim_{d(p,q)\to\infty} \frac{d(p,q)}{d_H(p,q)} = 1.$$

(As a matter of fact, the difference between d and d_H actually is uniformly bounded; i.e., there exists a constant C such that $|d(p,q) - d_H(p,q)| \le C$ for all p, q).

This provides us with valuable information about large-scale structure of left-invariant metrics on the Heisenberg group, for we have an explicit description of d_H. In particular, $d((0,0,0),(0,0,z))$ is approximately $\pi^{-1}\sqrt{|z|}$. We see that the z-axis does not look at all like a geodesic with respect to d—the distance between two far-away points in the z-axis is much smaller than the length of its segment.

Let us note that d_H possesses an additional symmetry as compared to d. Consider a transformation $A_t\colon \mathbb{R}^3 \to \mathbb{R}^3$ (t is a positive parameter) acting by $A_t(x,y,x) = (tx, ty, t^2 z)$. It follows from the following exercise that A_t is a homothety (dilation) with respect to d_H, that is, $d_H(A_t(p), A_t(q)) = t d_H(p,q)$.

Exercise 5.4.20. Check that A_t leaves the distribution H_p invariant, and it multiplies the length of curves given in (5.20) by t.

This implies that the (metric) tangent cone of $(\mathbb{R}^3, d_H)$ is isometric to $\mathbb{R}^3, d_H$ itself, and it in its turn is also isometric to its asymptotic cone (the cone at infinity). Using Exercise 5.4.19, we conclude that the asymptotic cone of the Heisenberg group with a left-invariant metric d is isometric to $(\mathbb{R}^3, d_H)$.

Information: The tangent cone of a Carnot–Carathéodory space at a generic point is isometric to a nilpotent group with a left-invariant sub-Riemannian metric; at "nongeneric" points it is isometric to a homogeneous space of a nilpotent group.

5.5. Riemannian and Finsler Volumes

In this section we take a more detailed and formal look at the notion of volume in Riemannian manifolds. We have already defined the area of a two-dimensional Riemannian region by an explicit formula (5.4). The motivation for this formula remains valid in any dimension. We will turn the motivation for this formula into a formal argument showing that the volume *must* equal a similar integral expression provided that it depends monotonically on the metric.

In other words, only one "reasonable" notion of Riemannian volume exists—no matter how one defines it, the result is the same. This is no longer true for Finsler metrics; as a result, there exist many different notions of Finsler volume. These issues are discussed in subsection 5.5.3.

We fix a dimension $n \geq 1$; all vector spaces and manifolds in this section are implicitly assumed to be n-dimensional.

5.5.1. Riemannian volume and Jacobians.

Definition 5.5.1. The *Riemannian volume* in an n-dimensional Riemannian manifold is the n-dimensional Hausdorff measure (cf. Section 1.7) determined by its Riemannian metric.

We denote the Riemannian volume by Vol. In case of ambiguity we add the manifold or Riemannian structure as an index (e.g., Vol_M).

In fact, we need only two properties of Riemannian volume:

(1) Riemannian volume in $\mathbb{R}^n$ is the standard Euclidean volume (the Lebesgue measure).

(2) Volume is monotone with respect to the metric. That is, if M and N are Riemannian manifolds and $f : M \to N$ is a distance-nonexpanding diffeomorphism, then $\mathrm{Vol}_N(f(\Omega)) \leq \mathrm{Vol}_M(\Omega)$ for any measurable set $\Omega \subset M$.

We will show that the Riemannian volume is uniquely determined by these properties. The purpose of the above definition was to make sure that such a volume functional exists.

According to the definition, the Riemannian volume is a Borel measure over a Riemannian manifold. Then one can use Lebesgue integration: if M is a Riemannian manifold and $f : \Omega \to \mathbb{R}$ is a nonnegative measurable (or, more generally, integrable) function on a measurable set $\Omega \subset M$, the integral $\int_\Omega f \, d\,\mathrm{Vol}_M$ is defined. We will use only elementary and natural properties of integration such as being linear and monotone with respect to the function being integrated. We refer to Chapter 2 of [**Fe**] for the general theory of measure and integration.

Jacobians. Let M and N be Riemannian manifolds of the same dimension n. Let $f : M \to N$ be a map differentiable at an $x \in M$. The derivative $d_x f$ is a linear map from $T_x M$ to $T_{f(x)} N$. Riemannian structures of M and N define Euclidean scalar products in $T_x M$ and $T_{f(x)} N$ making these two tangent spaces isometric to $\mathbb{R}^n$. In particular, $T_x M$ and $T_{f(x)} N$ are naturally equipped with n-dimensional Lebesgue (or Hausdorff) measure. Since every linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$ multiplies the volumes by a constant depending only on the map (see Exercise 1.7.6), so does the linear map $d_x f : T_x M \to T_{f(x)} N$. The respective constant is called the *Jacobian*:

Definition 5.5.2. Let M, N, f and x be as above. The *Jacobian* of f at x, denoted $\mathrm{Jac}\, f(x)$, is a real number such that

$$\mu_n(d_x f(X)) = \mathrm{Jac}\, f(x) \cdot \mu_n(X)$$

for all measurable sets $X \subset T_xM$.

If T_xM and $T_{f(x)}N$ are equipped with orthonormal bases and A is the matrix of d_xf in these bases, then $\operatorname{Jac} f(x) = |\det A|$ (cf. Exercise 1.7.6).

Remark 5.5.3. Our definition implies that the Jacobian is always nonnegative. There is a slightly different notion of Jacobian which is a *signed* quantity and equals the determinant of the corresponding matrix (not its absolute value). This is suitable for maps from $\mathbb{R}^n$ to $\mathbb{R}^n$ but the sign makes little sense for Jacobians in general (possibly nonorientable) Riemannian manifolds. We will never use signed Jacobians.

Jacobians determine how a map changes the Riemannian volume. The respective formula is known in analysis as the change-of-variable formula.

Theorem 5.5.4 (change of variable formula). *If M and N are Riemannian manifolds, $f : M \to N$ is a diffeomorphism, and $\Omega \subset M$ is a measurable set, then*

$$\operatorname{Vol}(f(\Omega)) = \int_\Omega \operatorname{Jac} f \, d\operatorname{Vol}_M .$$

In particular, $\operatorname{Vol}(N) = \int_M \operatorname{Jac} f$.

Proof. The formula is based on the following local statement: given $x \in M$, then for (small) measurable sets X containing x one has

$$\frac{\operatorname{Vol}_N(f(X))}{\operatorname{Vol}_M(X)} \to \operatorname{Jac} f(x) \qquad \text{as } \operatorname{diam}(X) \to 0. \tag{5.21}$$

To prove (5.21), fix a sufficiently small neighborhood U of x and a map $\varphi : U \to T_xM$ which is a diffeomorphisms from U to a neighborhood of 0 in T_xM such that $\varphi(x) = 0$ and $d_x\varphi : T_xM \to T_xM$ is the identity map. (For example, let φ be the inverse of $\exp_x$.) This map introduces local coordinates in U. Compare the Riemannian structure of M written in these coordinates with the Euclidean structure in T_xM. They coincide at the origin; hence their corresponding metrics are equal up to the first order near the origin. Or, equivalently, φ and φ^{-1} restricted to sufficiently small neighborhoods of 0 and x have Lipschitz constants arbitrarily close to 1. Since the volume is monotone with respect to metric, it follows that

$$\frac{\operatorname{Vol}_{T_xM}(\varphi(X))}{\operatorname{Vol}_M(X)} \to 1$$

as $\operatorname{diam}(X) \to 0$. Now let $y = f(x)$, $V = f(U)$ and a map $\psi : V \to T_yN$ be defined by the formula $\psi \circ f|U = d_xf \circ \varphi|U$. Then $\psi(y) = 0$ and $d_y\psi : T_yN \to T_yN$ is the identity. Similarly to the above, we obtain that

$$\frac{\operatorname{Vol}_{T_yN}(\psi(f(X)))}{\operatorname{Vol}_N(f(X))} \to 1$$

as $\operatorname{diam}(X) \to 0$. Since $\psi(f(X)) = d_x f(\varphi(X))$, the desired formula (5.21) now follows from the relation

$$\frac{\operatorname{Vol}_{T_y N}(d_x f(\varphi(X)))}{\operatorname{Vol}_{T_x M}(\varphi(X))} = \operatorname{Jac} f(x)$$

which is nothing but the definition of Jacobian.

How to derive the theorem from (5.21) depends on the reader's background. The most elementary method is to split X into small subsets $\{X_i\}$, choose points $x_i \in X_i$, and compare the volume $\operatorname{Vol}_N(f(X)) = \sum \operatorname{Vol}_N(f(X_i))$ with the integral sum $\sum \operatorname{Jac} f(x_i) \operatorname{Vol}_M(X_i)$. The details are left as an exercise. □

As a corollary to Theorem 5.5.4 we obtain the coordinate formula for the Riemannian volume which is often taken as the definition.

Theorem 5.5.5. *Let M be an n-dimensional Riemannian manifold, U an open set in $\mathbb{R}^n$ and $\varphi : U \to M$ a coordinate system. For an $x \in U$, let $(g_{ij}(x))$ be the matrix composed of the scalar products of the coordinate tangent vectors at $\varphi(x)$, that is, $g_{ij}(x) = \langle d_x\varphi(e_i), d_x\varphi(e_j)\rangle_M$ where $\{e_i\}$ is the standard basis of $\mathbb{R}^n$. Then*

$$\operatorname{Vol}_M(\varphi(\Omega)) = \int_\Omega \sqrt{\det(g_{ij})}\, dm_n$$

for any measurable $\Omega \subset U$.

Proof. Consider U as a Riemannian manifold whose scalar product is standard Euclidean. By Theorem 5.5.4, $\operatorname{Vol}_M(\varphi(U)) = \int_U \operatorname{Jac}\varphi$. It remains to prove that $\operatorname{Jac}\varphi(x) = \sqrt{\det(g_{ij}(x))}$ for every $x \in U$. Let (v_i) be an orthonormal basis of $T_\varphi(x)M$ and A be the matrix of $d_x\varphi : \mathbb{R}^n \to T_\varphi(x)M$ with respect to this basis. Then $\operatorname{Jac}\varphi(x) = |\det A|$. On the other hand, $(g_{ij}(x)) = A^T A$ where A^T is the transpose matrix. Therefore $\det(g_{ij}(x)) = \det(A^T)\det(A) = \det(A)^2$; hence $|\det A| = \sqrt{\det(g_{ij}(x))}$. The theorem follows. □

Note that Theorem 5.5.5 uniquely defines the volume of any measurable set in a Riemannian manifold. On the other hand, in its proof (including the proof of the change of variable formula) we only used the two basic properties mentioned after Definition 5.5.1. Hence these two properties uniquely determine the volume, and we obtain the following

Theorem 5.5.6. *Let V be a function associating to every Riemannian manifold M a Borel measure over it. Suppose that V is monotone with respect to the metric and yields the standard Euclidean volume in $\mathbb{R}^n$. Then V coincides with the Riemannian volume.*

5.5.2. Volume of Lipschitz maps. Though we proved the change of variable formula only for diffeomorphisms, it works for arbitrary Lipschitz homeomorphisms. Moreover, a simple modification makes it valid for Lipschitz maps that are not bijective. Below we give the respective formulations without proofs.

Theorem 5.5.7 (Rademacher's theorem; see Theorem 3.1.6 in [**Fe**]). *Every Lipschitz map is differentiable almost everywhere. That is, if M and N are Riemannian manifolds and $f : M \to N$ is a Lipschitz map, then the derivative $d_x f$ exists for all $x \in M$ except a set of zero measure.*

Note that the word "Riemannian" in this theorem can be replaced by "smooth" because the class of Lipschitz maps and the class of sets of zero measure does not depend on the choice of Riemannian structure (in fact, both notions can be defined in terms of local coordinates).

Rademacher's theorem implies that for a Lipschitz map $f : M \to N$ the Jacobian $\operatorname{Jac} f(x)$ is defined for almost all $x \in M$. Therefore the Lebesgue integral $\int_M \operatorname{Jac} f$ is defined. This integral is called the volume (or area) of f. If f is injective, it equals the volume of $f(M)$ in N. If f covers some parts of N multiple times, the multiplicity should be taken into account:

Theorem 5.5.8 ([**Fe**], Theorem 3.2.3). *If M and N are n-dimensional Riemannian manifolds and $f : M \to N$ is a Lipschitz map, then*

$$\int_M \operatorname{Jac} f(x)\, d\operatorname{Vol}_M(x) = \int_N \#(f^{-1}(y))\, d\operatorname{Vol}_N(y)$$

where $\#$ denotes the cardinality.

The integral on the right-hand side is, of course, the Lebesgue integral of the function $y \mapsto \#(f^{-1}(y))$ which is always measurable. Since the function takes only integer values and the value infinity, its integral can be written as

$$\int_N \#(f^{-1}(y))\, d\operatorname{Vol}_M(y) = \sum_{k \in \mathbb{N} \cup \{\infty\}} k \cdot \operatorname{Vol}_N(\{y \in N : \#(f^{-1}(y)) = k\})$$

where the term $\infty \cdot 0$, if it appears, equals 0.

One of the most common applications of the above theorem is the following upper bound for the volume of a map's image.

Corollary 5.5.9. *If M and N are n-dimensional Riemannian manifolds and $f : M \to N$ is a Lipschitz map, then*

$$\operatorname{Vol}_N(f(M)) \le \int_M \operatorname{Jac} f(x)\, d\operatorname{Vol}_M(x).$$

In particular, if $\operatorname{Jac} f \le 1$ almost everywhere, then $\operatorname{Vol}_N(f(M)) \le \operatorname{Vol}(M)$.

Proof. $\#(f^{-1}(y)) \geq 1$ if $y \in f(M)$. Substitute this into Theorem 5.5.8. □

The next proposition is used in Section 5.6.

Proposition 5.5.10. *Let M be an n-dimensional Riemannian manifold, $f : M \to \mathbb{R}^n$ be a Lipschitz map and f_i ($1 \leq i \leq n$) be its coordinate functions. Then* $\operatorname{Jac} f \leq \prod_{i=1}^m \operatorname{dil}(f_i)$ *wherever f is differentiable.*

Proof. Let f be differentiable at a point $x \in M$. It is easy to see that $|d_x f_i(v)| \leq \operatorname{dil}(f_i) \cdot |v|$ for all $v \in T_x M$, i.e., $\|d_x f_i\| \leq \operatorname{dil}(f_i)$. Let $A = (a_{ij})$ be a matrix of $d_x f$ with respect to an orthonormal basis of $T_x M$ and the standard basis of $\mathbb{R}^n$. The ith row of this matrix is composed of the numbers $d_x f_i(v_j)$, $1 \leq j \leq n$. Let us interpret this row as a vector $a_i \in \mathbb{R}^n$; then $|a_i| = \|d_x f_i\| \leq \operatorname{dil}(f_i)$. Since $|\det A|$ equals the Euclidean volume of the parallelotope spanned by the vectors a_i, we have $\operatorname{Jac} f(x) = |\det A| \leq \prod_{i=1}^n |a_i| \leq \prod_{i=1}^n \operatorname{dil}(f_i)$. □

5.5.3. Finslerian volumes. Theorem 5.5.6 tells us that there is only one reasonable notion of volume for Riemannian manifolds. This is not the case with Finsler metrics. One can define Finslerian volume in different ways and obtain essentially different results. Some examples are given below in this section.

For now, we assume that some Finslerian volume functional is fixed. That is, every Finsler manifold is equipped with a Borel measure (depending on its Finsler structure) that we denote by Vol. We require that Vol satisfy the properties listed after Definition 5.5.1, namely, the Euclidean compatibility and monotonicity with respect to metric. Note that monotonicity implies that (smooth) isometries preserve the volume.

Let $\|\cdot\|$ be a norm in $\mathbb{R}^n$. As a particular case of a Finsler manifold, the normed space $(\mathbb{R}^n, \|\cdot\|)$ carries a Finslerian volume, $\mathrm{Vol}_{\|\cdot\|}$. Let $|\cdot|$ be the Euclidean norm. Since $\|\cdot\|$ and $|\cdot|$ are bi-Lipschitz equivalent (Theorem 1.4.11), there exist positive constants c and C such that $c|x| \leq \|x\| \leq C|x|$ for all $x \in \mathbb{R}^n$. The monotonicity of volume then implies that $c^n m_n \leq \mathrm{Vol}_{\|\cdot\|} \leq C^n m_n$. (To prove this, consider the identity map as a map from $(\mathbb{R}^n, \|\cdot\|)$ to $(\mathbb{R}^n, c|\cdot|)$ and from $(\mathbb{R}^n, C|\cdot|)$ to $(\mathbb{R}^n, \|\cdot\|)$.) In particular, the $\|\cdot\|$-volume of a unit cube is finite and positive. We denote this quantity $\mathrm{Vol}_{\|\cdot\|}([0,1]^n)$ by $\nu(\|\cdot\|)$.

The volume $\mathrm{Vol}_{\|\cdot\|}$ is preserved by parallel translations since they are isometries of $(\mathbb{R}^n, \|\cdot\|)$. By Lebesgue's theorem (1.7.5) it follows that the measure $\frac{\mathrm{Vol}_{\|\cdot\|}}{\nu(\|\cdot\|)}$ coincides with the Lebesgue measure m_n. Thus

$$\mathrm{Vol}_{\|\cdot\|}(X) = \nu(\|\cdot\|)\, m_n(X)$$

for any measurable set $X \subset \mathbb{R}^n$.

Recall that a Finsler structure Φ in a region $U \subset \mathbb{R}^n$ is a continuous function on TU whose restriction to every tangent space T_xU (where $x \in U$) is a vector-space norm. We denote this restriction by Φ_x.

Proposition 5.5.11. *If Φ is a Finsler structure in a region $U \subset \mathbb{R}^n$, then*

$$\mathrm{Vol}_\Phi(\Omega) = \int_\Omega \nu(\Phi_x)\, dm_n(x)$$

for any measurable set $\Omega \subset U$.

Proof. Similar to Theorems 5.5.4 and 5.5.5. □

This proposition provides a sort of "general form" of a Finslerian volume. One can define a Finslerian volume functional by specifying a value $\nu(\|\cdot\|)$ for every norm $\|\cdot\|$ in $\mathbb{R}^n$. In other words, an n-dimensional Finslerian volume is determined by its values on flat normed spaces. More precisely, the following proposition holds.

Proposition 5.5.12. *Suppose that every normed space $(V, \|\cdot\|)$ is equipped with a translation-invariant measure $\mathrm{Vol}_{(V,\|\cdot\|)}$ which is finite and positive on open bounded sets, so that the following conditions are satisfied.*

(1) *Euclidean compatibility:* $\mathrm{Vol}_{(\mathbb{R}^n, \textit{standard Euclidean norm})} = m_n$.
(2) *Affine invariance: any linear isometry between normed spaces preserves the measure.*
(3) *Monotonicity: if $\|\cdot\| \le \|\cdot\|'$, then* $\mathrm{Vol}_{(V,\|\cdot\|)} \le \mathrm{Vol}_{(V,\|\cdot\|')}$.

Then this family of measures can be extended to all Finsler manifolds as a Euclidean-compatible and monotone volume functional.

A plan of proof. Having a measure over every normed space allows one to define the Jacobian of a map from one Finsler manifold to another, similarly to Definition 5.5.2. Define Finslerian volume by the formula from Proposition 5.5.11; then the change of variable formula (Theorem 5.5.4) follows easily. The conditions 2 and 3 imply that a linear nonexpanding map does not increase the measure. Observe that derivatives of a nonexpanding maps between Finsler manifolds are nonexpanding linear maps between their tangent spaces (equipped with norms restricted from the Finslerian structures). Then the change of variable formula implies that the Finslerian volume is well-defined (i.e., does not depend on the choice of a coordinate system) and monotone with respect to metric. □

In order to define a translation-invariant measure over a vector space, it is sufficient to specify its value on a single set. Above we used the unit cube in $\mathbb{R}^n$ for this purpose (recall the definition of $\nu(\|\cdot\|)$). In most cases it is

more convenient to fix a volume of a set which is naturally associated with the norm, for example, the norm's unit ball.

Example 5.5.13 (Hausdorff measure). Let α_n denote the Euclidean volume of the standard unit ball in $\mathbb{R}^n$. For every normed space $(V, \|\cdot\|)$ define a measure $\mathrm{Vol}_{(V,\|\cdot\|)}$ so that the measure of the norm's unit ball equals α_n. Equivalently, define

$$\nu(\|\cdot\|) = \frac{\alpha_n}{m_n(\text{unit ball of } \|\cdot\|)}$$

for a norm $\|\cdot\|$ in $\mathbb{R}^n$. It is easy to see that the conditions of Proposition 5.5.12 are satisfied, so we obtain some Finsler volume functional. In fact, this functional coincides with the n-dimensional Hausdorff measure. The proof of this coincidence is similar to determining the precise value of the normalization constant of the Hausdorff measure (see Theorem 1.7.14 and its corollary).

The definition of volume in the following example requires choosing a set of minimal volume from a certain class of sets in a vector space. This looks like a vicious circle because the volume itself is not yet defined. However, one does not need to fix a volume functional to *compare* volumes of sets. Fix an arbitrary Euclidean structure and compare the Lebesgue measures; the result does not depend on that auxiliary structure because all measures are the same up to a multiplicative constant.

Example 5.5.14 (comass, [**Gro1**]). Let $(V, \|\cdot\|)$ be a normed space, B its unit ball, and Q is an affine cube of minimal volume containing B. (By affine cube we mean an image of the cube $[-1,1]^n$ under a linear map.) Then define the volume by the formula $\mathrm{Vol}_{(V,\|\cdot\|)}(Q) = 2^n$. This defines the Finslerian volume functional called *comass*.

Example 5.5.15 (inscribed Riemannian volume). Similarly to the above, let Q be an ellipsoid of maximal volume contained in B. Let the volume $\mathrm{Vol}_{(V,\|\cdot\|)}(Q)$ equal α_n, the Euclidean volume of the standard Euclidean ball. An equivalent definition is: for a given norm $\|\cdot\|$, find a Euclidean norm $|\cdot|$ whose volume form is minimal among all Euclidean norms which are greater or equal to $\|\cdot\|$; then set $\mathrm{Vol}_{(V,\|\cdot\|)} = \mathrm{Vol}_{(V,|\cdot|)}$. The resulting Finslerian volume equals the minimal Riemannian volume of a Riemannian metric which is greater or equal to the given Finsler one.

Similarly, one can consider the maximal Riemannian volume of a Riemannian metric which is less than or equal to a given Finsler one. This definition of volume is related to the minimal ellipsoid containing the norm's unit ball.

Example 5.5.16 (symplectic volume). Let $\|\cdot\|$ be a norm in $\mathbb{R}^n$ and B its unit ball. A *polar* set to B is defined by

$$B^* = \{x \in \mathbb{R}^n : \langle x, y\rangle \leq 1 \text{ for all } y \in B\}.$$

Define the volume functional by $\nu(\|\cdot\|) = m_n(B^*)/\alpha_n$ where α_n is the Euclidean volume of the Euclidean unit ball (compare Example 5.5.13). The resulting Finsler volume functional is called the *Holmes–Thompson volume.*

This particular Finsler volume is related to the so-called symplectic volume (which is defined over the co-tangent space of the manifold and does not depend on the metric). Precise formulation: the Holmes–Thompson volume equals a constant multiplied by the projection of the symplectic volume from the set of co-vectors having less than unit norm.

Exercise 5.5.17. Prove that all Finsler volume functionals in the above examples are mutually different.

While there are many different natural Finsler volume functionals, they are nevertheless "not too different". Namely the ratio of any two of them is bounded by a constant depending only on dimension:

Theorem 5.5.18. *Let* Vol *and* Vol′ *be two n-dimensional Euclidean-compatible monotone Finsler volume functionals. Then* $\mathrm{Vol}(\Omega) \leq n^{3n/2}\,\mathrm{Vol}'(\Omega)$ *for any measurable set* Ω *is any Finsler manifold.*

Proof. By Proposition 5.5.11, it is sufficient to prove the inequality $\mathrm{Vol} \leq n^{3n/2}\,\mathrm{Vol}'$ in normed vector spaces. We need the following lemma (also used in Subsection 8.5.3). Roughly speaking, it claims that every symmetric convex body in $\mathbb{R}^n$ can be approximated, with a bounded relative precision, by an affine cube. Recall that an affine cube is an image of the cube $[-1,1]^n$ under a nondegenerated linear map.

Lemma 5.5.19. *Let D be a unit ball of a norm $\|\cdot\|$ in an n-dimensional vector space V. Then there exists an affine cube $Q \subset V$ such that $\frac{1}{n}Q \subset D \subset Q$.*

Proof. We may assume that $V = \mathbb{R}^n$. Consider the function F on $D \times \dots \times D$ (n times) defined by $F(v_1, \dots, v_n) = |\det[v_1, \dots, v_n]|$ where $[v_1, \dots, v_n]$ is the $n \times n$ matrix composed of the coordinates of the vectors v_i. In other words, $F(v_1, \dots, v_n)$ is the volume of the parallelotope spanned by these vectors. Since F is continuous and $D \times \dots \times D$ is compact, F attains a maximum. We assume that this maximum is achieved on the standard basis $(e_1, \dots, e_n)$ of $\mathbb{R}^n$. This can be achieved by applying a linear transformation to D. Then one can let $Q = [-1,1]^n$. Indeed, since D contains the vectors e_i, one has $\|e_i\| \leq 1$; hence $\|(x_1, \dots, x_n)\| = \|\sum x_i e_i\| \leq \sum |x_i| \leq 1$

whenever $(x_1, \dots, x_n) \in [-\frac{1}{n}, \frac{1}{n}]^n$. Therefore $\frac{1}{n}Q \subset D$. On the other hand, if $v = (x_1, \dots, x_n) \in D$, then $|x_i| \le 1$ for all i; otherwise one could replace e_i by v in the expression $F(e_1, \dots, e_n)$ and obtain a greater value. Thus $D \subset [-1, 1]^n$. □

Now let $(V, \|\cdot\|)$ be an n-dimensional normed space. Let Q be an affine cube from Lemma 5.5.19. Identify V with $\mathbb{R}^n$ by means of a linear isomorphism so that $Q = [-1, 1]^n$. Then the relation $\frac{1}{n}Q \subset D \subset Q$ implies that $\frac{1}{n}B \subset D \subset \sqrt{n}B$ where B is the Euclidean unit ball in $\mathbb{R}^n$. This is in turn equivalent to an inequality between norms: $\frac{1}{\sqrt{n}}|\cdot| \le \|\cdot\| \le n|\cdot|$, where $|\cdot|$ is the standard Euclidean norm. Hence

$$\frac{1}{n^{n/2}} \operatorname{Vol}_{(\mathbb{R}^n, |\cdot|)} \le \operatorname{Vol}_{(\mathbb{R}^n, \|\cdot\|)} \le n^n \operatorname{Vol}_{(\mathbb{R}^n, |\cdot|)}$$

by monotonicity of volume (note that $\operatorname{Vol}_{(\mathbb{R}^n, c|\cdot|)} = c^n \operatorname{Vol}_{(\mathbb{R}^n, |\cdot|)}$ for any constant $c > 0$). Similarly, the same inequalities hold for Vol' in place of Vol. Since $\operatorname{Vol}_{(\mathbb{R}^n, |\cdot|)} = \operatorname{Vol}'_{(\mathbb{R}^n, |\cdot|)}$, the desired inequality $\operatorname{Vol} \le n^{3n/2} \operatorname{Vol}'$ follows. □

5.6. Besikovitch Inequality

The Besikovitch inequality is one of the simplest facts that allows one to estimate the volume of a Riemannian metric if certain (very limited) information about distances is known. It belongs to the "curvature-free" part of metric geometry in the sense that its formulation does not include curvature bounds or similar local conditions. Many other results and problems in this area can be found in [**Gro1**]. We formulate and prove the Besikovitch inequality in subsection 5.6.2. The proof relies on the theory of degree (which is a part of differential topology) that we briefly discuss in the first subsection. While this theory is far from the topic of this book, it provides important tools commonly used in metric geometry.

5.6.1. Degree of a map. In this subsection we introduce an important homotopy invariant of a map, the degree. We mainly discuss the degree modulo 2 which is sufficient for our purposes. For details and proofs see [**Mi**]. We recommend the reader prove all the statements below as exercises in the one-dimensional case (i.e., for circles or intervals).

Definition 5.6.1. Let M and N be smooth manifolds and $f : M \to N$ a smooth map. An $x \in M$ is called a *regular point* for f if the rank of $d_x f$ equals the dimension of N, i.e., $d_x f : T_x M \to T_{f(x)} N$ is surjective. A point $y \in N$ is called a *regular value* of f if every point $x \in f^{-1}(y)$ is a regular point for f.

The well-known Sard–Brown Theorem says that regular values of f are dense in N (moreover, the complement of the set of regular values in N is a set of zero measure).

Note that any point $y \in N$ such that $f^{-1}(y) = \emptyset$ is a regular value. If $\dim M < \dim N$ the set of regular values coincides with $N \setminus f(M)$. The definition is more interesting if $\dim M \geq \dim N$. In this case the implicit function theorem implies that, if y is a regular value of f, then $f^{-1}(y)$ is a smooth submanifold of M whose dimension equals $\dim M - \dim N$. In particular, in the case $\dim M = \dim N$ the inverse image of a regular value is discrete (consists of isolated points). If M is compact, the inverse image of a regular value is a finite set.

The theory of degree applies only to maps between manifolds of equal dimensions. Below M and N always denote compact smooth manifolds of dimension $n \geq 1$.

Definition 5.6.2. Let $f : M \to N$ be a smooth map and $y \in N$ be a regular value of f. Then the quantity $\deg_2(f; y)$ is a residue modulo 2 defined by

$$\deg_2(f; y) = \#(f^{-1}(y)) \bmod 2.$$

In other words $\deg_2(f; y)$ equals 0 if the number of points in $f^{-1}(y)$ is even, and 1 otherwise.

Proposition 5.6.3. $\deg_2(f; y)$ *does not depend on the choice of a regular value* $y \in N$.

This proposition, along with the fact that at least one regular value exists, allows us to introduce the following

Definition 5.6.4. Let $f : M \to N$ be a smooth map. Then $\deg_2(f; y)$, where $y \in N$ is an arbitrary regular value of f, is called the *degree modulo 2* of f and denoted $\deg_2(f)$.

Proposition 5.6.5. *If smooth maps* f_1 *and* f_2 *from* M *to* N *are homotopic, then* $\deg_2(f_1) = \deg_2(f_2)$.

This proposition allows us to apply the notion of degree to nonsmooth maps. It can be shown that every continuous map from M to N is homotopic to a smooth map. If $f : M \to N$ is an arbitrary continuous map and f_1 is a smooth map which is homotopic to f, one can define $\deg_2(f) = \deg_2(f_1)$. Proposition 5.6.5 implies that $\deg_2(f_1)$ does not depend on the choice of f_1, so $\deg_2(f)$ is well-defined by this formula. Moreover, $\deg_2(f)$ depend only on the homotopy class of f.

Manifolds with boundary. In the above considerations we assumed that M and N have no boundary. This assumption can be dropped if one restricts

the class of maps to ones that map the boundary of M to (a subset of) the boundary of N.

Here by a manifold with boundary we mean a smooth manifold with a piecewise-smooth boundary. In fact, the degree can be defined without any differential structure (i.e., it applies to topological manifolds).

Proposition 5.6.6. *Suppose that M and N are compact smooth manifolds, possibly with boundaries, and f is a smooth map such that $f(\partial M) \subset \partial N$. Then $\deg_2(f;y)$ does not depend on a regular value $y \in N$. Hence Definition 5.6.4 applies to such a map f.*

Furthermore, if f_0 and f_1 are connected via a homotopy $\{f_t\}_{t\in[0,1]}$ such that $f_t(\partial M) \subset \partial N$ for all $t \in [0,1]$, then $\deg_2(f_0) = \deg_2(f_1)$.

Proof. We reduce the statements to the case of closed manifolds (Propositions 5.6.3 and 5.6.5) by means of the doubling construction. Let $\bar{M}$ be the double of M, that is, $\bar{M}$ consists of two copies of M glued along the boundary, and $\bar{N}$ is the double of N. Then a map $f : M \to N$ naturally defines a map $\bar{f} : \bar{M} \to \bar{N}$ which maps each half-manifold of $\bar{M}$ to the respective half-manifold of $\bar{N}$ in the same way as f does. If $f(\partial M) \subset \partial N$, the new map $\bar{f}$ is continuous and it is easy to see that $\deg_2(f;y) = \deg_2(\bar{f};y')$ where y' is one of the two points corresponding to y. Then Propositions 5.6.3 implies the first statement. Similarly, Proposition 5.6.5 implies the second one. The details (including the issue of introducing a differential structure on the doubles) are left to the reader. □

As we mentioned after Proposition 5.6.5, the homotopy invariance of $\deg_2$ allows us to define $\deg_2$ for nonsmooth maps. The same argument applies to manifolds with boundaries and the class of maps f such that $f(\partial M) \subset \partial N$.

Proposition 5.6.7. *If M and N are compact smooth n-manifolds, possibly with boundaries, $f : M \to N$ is a continuous map such that $f(\partial M) \subset \partial N$ and $\deg_2 f \neq 0$, then f is surjective, i.e., $f(M) = N$.*

Proof. First, reduce the proposition to the case of no boundary just like Proposition 5.6.6. Then suppose f is not surjective. If f is smooth, consider a $y \in N \setminus f(M)$. Since $f^{-1}(y) = \emptyset$, y is a regular value and $\deg_2(f) = \deg_2(f;y) = 0$. In the general case, replace f by a smooth approximation f_1. If f_1 is sufficiently close to f, it is homotopic to f and still nonsurjective; hence $\deg_2(f) = \deg_2(f_1) = 0$. □

Remark 5.6.8. If M and N are *oriented* manifolds, one can define the integer-valued degree of a map $f : M \to N$, denoted $\deg(f)$. Namely, if

$y \in N$ is a regular value of f, let

$$\deg(f; y) = \sum_{x \in f^{-1}(y)} \varepsilon(x)$$

where

$$\varepsilon(x) = \begin{cases} +1, & \text{if } d_x f \text{ is orientation-preserving,} \\ -1, & \text{otherwise.} \end{cases}$$

All properties of $\deg_2$ that we formulated hold for deg with obvious modifications. It is clear that $\deg_2(f) = \deg(f) \bmod 2$.

5.6.2. Besikovitch inequality. We will use the following notation. Let $I = [0,1]$; then $I^n = [0,1]^n \subset \mathbb{R}^n$ is the standard n-dimensional cube. The boundary ∂I^n consists of points where at least one of the coordinates equals 0 or 1. This boundary is the union of *faces* that we denote F_i^0 and F_i^1; namely, F_i^0 (resp. F_i^1) is the set of points in I^n whose ith coordinate equals 0 (resp. 1).

Theorem 5.6.9 (Besikovitch inequality). *Let g be a Riemannian structure in I^n. For $i = 1, \dots, n$ let d_i denote the Riemannian distance in this metric between the faces F_i^0 and F_i^1. Then* $\mathrm{Vol}_g(I^n) \geq \prod_{i=1}^n d_i$.

Proof. During this proof all distances, Jacobians, etc., on I^n are taken with respect to g. For each $i = 1, \dots, n$ define a function $f_i : I^n \to \mathbb{R}$ by

$$f_i(x) = \min\{d_i, \mathrm{dist}(x, F_i^0)\}$$

and let $f : I^n \to \mathbb{R}^n$ be the function whose coordinate functions are f_i, i.e., $f(x) = (f_1(x), \dots, f_n(x))$. Observe the following trivial facts.

(1) The functions f_i are nonexpanding.
(2) $f_i(x) \in [0, d_i]$ for all $x \in I^n$.
(3) $f_i(x) = 0$ if $x \in F_i^0$ and $f_i(x) = d_i$ if $x \in F_i^1$. In other words, f maps each face of I^n to the corresponding face of the parallelotope P.

By Proposition 5.5.10 the first of the above facts implies that $\mathrm{Jac}\, f \leq 1$ almost everywhere. By Corollary 5.5.9 it follows that $\mathrm{Vol}_g(I^n) \geq \mu_n(f(I^n))$. We will show that $f(I^n) = P$.

The second of the above facts allows us to consider f as a map from I^n to P. The third one implies that $f(\partial I^n) \subset \partial P$; hence the notion of degree applies. We will show that the degree of f (modulo 2) as a map from I^n to P equals 1. Consider the linear map $A : \mathbb{R}^n \to \mathbb{R}^n$ defined by $A(x_1, \dots, x_n) = (x_1/d_1, \dots, x_n/d_n)$. Its restriction $A|_P$ is a homeomorphism from P to I^n. Hence it is sufficient to prove that $A \circ f$ has degree 1 as a map from I^n to I^n (then apply Proposition 5.6.7). We will prove this by

showing that $A \circ f$ is homotopic to the identity via a homotopy $\{\varphi_t\}_{t\in[0,1]}$ such that $\varphi_t(\partial I^n) \subset \partial I^n$ for all t (see Proposition 5.6.6). Let $\{\varphi_t\}$ be a linear homotopy between $A\circ f$ and the identity, i.e., $\varphi_t(x) = (1-t)A(f(x))+tx$. Let $x \in \partial I^n$ and let F be a face of I^n to which x belongs. Then $y := A(f(x)) \in F$ since $A \circ f|_{I^n}$ maps each face of I^n to itself. It follows immediately that $\varphi_t(x) = tx + (1-t)y \in F \subset \partial I^n$ for all t.

Thus $\varphi_t(\partial I^n) \subset \partial I^n$ for all t. Therefore $\deg_2(f) = \deg_2(A \circ f) = \deg_2(Id_{I^n}) = 1$ as we claimed above. In particular, the image of f is the entire P. Then

$$\mathrm{Vol}_g(I^n) \geq \mathrm{Vol}(f(I^n)) = \mathrm{Vol}(P) = \prod_{i=1}^{n} d_i.$$

□

Exercise 5.6.10. Prove the following generalizations of Theorem 5.6.9.

1. Let M be an n-dimensional Riemannian manifold with $\partial M = \partial I^n$. Prove that $\mathrm{Vol}(M) \geq \prod_{i=1}^n \mathrm{dist}_M(F_i^0, F_i^1)$.

2. Let M be an n-dimensional Riemannian manifold and $f : \partial M \to \partial I^n$ be a continuous map with nonzero degree modulo 2. Then $\mathrm{Vol}(M) \geq \prod_{i=1}^n \mathrm{dist}_M(f^{-1}(F_i^0), f^{-1}(F_i^1))$.

Exercise 5.6.11. Let M be a Riemannian manifold homeomorphic to the sphere S^2. Suppose that there are four points $a, b, c, d \in M$ with the following distances between them: $|ab| = |bc| = |cd| = |da| = 1$, $|ac| = |bd| = 3/2$. Prove that the area of M is at least $1/2$.

Hint: Prove that the distance between a shortest path from a to b and one from c to d is not less than $1/2$, and similarly for shortest paths from b to c and from d to a. Then apply the Besikovitch inequality to each of the two quadrilaterals into which these four paths divide the sphere.

Exercise 5.6.12. Let M be a Riemannian manifold homeomorphic to the projective plane $\mathbf{RP}^2$. Suppose that the length of any noncontractible loop in M is at least 1. Prove that the area of M is at least $1/16$.

Hint: Cut M along a shortest noncontractible loop. The resulting space is homeomorphic to a disc. Divide its boundary into four arcs of equal lengths and prove that the distance between "opposite" arcs is not less than $1/4$. Then apply the Besikovitch inequality to this disc (which is homeomorphic to I^2).

Remark 5.6.13. In fact, the constant $1/16$ above can be replaced by $1/4\pi$ (this is known as the theorem of Pu, see [**Gro1**]). The constant $1/4\pi$ is optimal and is achieved for the metric on $\mathbf{RP}^2$ space obtained from the sphere of radius $1/2\pi$ (by identifying opposite points).

Exercise 5.6.14. Let M be a Riemannian manifold homeomorphic to the torus $T^2 = S^1 \times S^1$. Suppose that the length of any noncontractible loop in M is at least 1. Prove that the area of M is at least $1/4$.

Remark 5.6.15. The optimal constant in the statement of the above exercise is $\sqrt{3}/2$ (Loewner's theorem; see [**Gro1**]). Can you find an example for which this is achieved?

Finally, let us mention that there is no "inverse Besikovitch inequality" (i.e., no similar *upper* bound for the volume):

Exercise 5.6.16. Prove that for every $C > 0$ and every integer $n \geq 2$ there exists a Riemannian structure g in the n-cube I^n such that the Hausdorff distance (in (I^n, dist_g)) between F_i^0 and F_i^1 is less than 1 for all i, but $\mathrm{Vol}_g(I^n) > C$.

5.6.3. Generalization: systoles. Exercises 5.6.12 and 5.6.14 are particular cases of a general problem which can be formulated as follows.

Let (M, g) be a compact n-dimensional Riemannian manifold. Its *one-dimensional systole*, denoted by $\mathrm{sys}_1(M, g)$, is the infimum of lengths of noncontractible loops in M. (The definition of k-dimensional systoles for $k > 1$ is more technical; see below.) The problem is: given the topological type of M and the value of $\mathrm{sys}_1(M, g)$, can one estimate the volume of (M, g) from below? For example, if M is homeomorphic to the torus T^2 or the projective plane $\mathbf{RP}^2$, one can prove that $\mathrm{Vol}(M, g) \geq \frac{\sqrt{3}}{2} \mathrm{sys}_1(M, g)^2$ and $\mathrm{Vol}(M, g) \geq \frac{1}{4\pi} \mathrm{sys}_1(M, g)^2$, respectively (see remarks after Exercises 5.6.12 and 5.6.14).

More formally, for an M, one defines the *isosystolic constant* $c_1(M)$ of an n-dimensional manifold M by

$$c_1(M) = \inf_g \{\mathrm{Vol}(M, g) : \mathrm{sys}_1(M, g) \geq 1\},$$

or, equivalently,

$$c_1(M) = \inf_g \frac{\mathrm{Vol}(M, g)}{\mathrm{sys}_1(M, g)^n}$$

where the infimum is taken over all Riemannian metrics g on M. One then asks, for a given topological type of M, what is the value of this constant, or at least whether it is positive or not.

Though the precise values of $c_1(M)$ are currently known only in a few simplest cases (all of which are two-dimensional), a vast class of manifolds M with $c_1(M) > 0$ has been found (see [**Gro1**]).

The notion of systole can be generalized to higher dimensions in a variety of ways. One possible approach is that the k-dimensional systole of a Riemannian manifold M is the infimum of k-volumes of homologically

nontrivial closed k-dimensional "films" in M. Here one has to specify what is meant by a "film". It is not good to consider only k-submanifolds as films; one of the reasons is that not all homology classes can be represented by submanifolds. Instead, one takes the infimum over all singular Lipschitz chains.

Let us explain this in more detail. Let τ be a k-dimensional Lipschitz singular simplex, that is, a Lipschitz map of the standard k-simplex into M. For such a simplex, the k-dimensional volume $\mathrm{Vol}(\tau)$ is well-defined (see Section 5.5.2). This allows one to introduce the notion of volume for k-dimensional Lipschitz chains. Namely for $c = \sum_{i=1}^{N} x_i \tau_i$, where $x_i \in \mathbb{R}$ and τ_i are Lipschitz singular simplices, define $\mathrm{Vol}(c) = \sum_{i=1}^{N} |x_i| \mathrm{Vol}(\tau_i)$. The boundary of a chain, cycles and homology groups are defined as usual. Finally one defines

$$\mathrm{sys}_k(M, g) = \inf\{\mathrm{Vol}(c) : \ c \text{ is a } k\text{-dimensional cycle not homological to } 0\}$$

and

$$c_k(M) = \inf_g \frac{\mathrm{Vol}(M, g)}{\mathrm{sys}_k(M, g)^{n/k}},$$

and asks whether $c_k(M)$ is positive or not.

It is a bit surprising that the answer is absolutely different for $k = 1$ and $k > 1$. While $c_1(M) > 0$ in many cases, it has been proved recently ([**KS**] and [**Bab**]) that one always has $\mathrm{sys}_k(M) = 0$ for $1 < k < \dim M$. This property is called *systolic softness*. Actually in [**Bab**] a more general fact on so-called *inter-systoles* is proved; furthermore, M is allowed to be a polyhedral space instead of a manifold.

Chapter 6

Curvature of Riemannian Metrics

The main conclusion of this chapter is the fact that a Riemannian manifold of sectional curvature bounded by k is a length space with the same curvature bound in the sense of Chapter 4. Hence the reader who is ready to take this statement on faith (or who is familiar with Riemannian geometry) can omit this chapter. It is included mainly to make this book more self-contained.

In our introduction to Riemannian geometry we basically stop where substantial Riemannian geometry begins. There are many textbooks containing various accounts of methods of differential and Riemannian geometry. In this course we regard Riemannian manifolds (as well as other length spaces whose definitions rely on calculus of variations) just as one of the important (or even leading) sources of examples. From this viewpoint, all differential methods can retire after they produce certain basic information to feed into the synthetic machinery. A shortest path leading from infinitesimal definitions to local metric properties is surprisingly short, and we intentionally make it somewhat longer to provide better motivations. Basically all we need at the end is to convert a Riemannian metric to normal coordinates, and then prove the (local) Cartan–Alexandrov–Toponogov Comparison Theorems 6.5.6.

For the sake of simplicity of exposition we restrict our introduction to Riemannian geometry mainly to the case of two-dimensional manifolds; this also allows us to give explicit formulas avoiding cumbersome indices and to eliminate linear algebra. Higher-dimensional generalizations are mostly straightforward and left to the reader (as an important exercise). We point

out a construction where passing to the higher-dimensional case may cause some difficulty and indicate how to overcome it.

One can introduce sectional (Gaussian) curvature, as well as understand its geometric meaning, even without defining covariant differentiation at all. To use this notion introduced this way one would have, however, to take a certain (rather technical) statement on faith. Recall that the cornerstone notion in the previous chapter was normal coordinates, and we had two (dual) approaches to them. If one begins with an equidistant family of curves (propagation of wavefronts), the geodesic curvature of a curve in this family is defined as the derivative of its length element. Then the Gaussian curvature governs the derivative of this geodesic curvature across the equidistant family. If one thinks of normal coordinates as a family of rays (geodesics), then their divergence is governed by the Gaussian curvature. In the two cases, Gaussian curvature enters as a coefficient in certain very simple differential equations (first-order and second-order, respectively). This coefficient is independent of the choice of a particular normal coordinate system (and depends on a point only, or on a point and a two-dimensional direction in the higher-dimensional case)—and we use the covariant derivative just to prove this fact (of course, it also can be obtained by quite cumbersome coordinate computations).

Since our considerations are local, we deal with a metric defined in a region. Of course a reader familiar with smooth manifolds will be able to immediately reformulate our statements for Riemannian metrics on smooth manifolds.

Here is a brief plan of the chapter.

First we begin with a coordinate computation. Using a direct variational argument, we show that every shortest path satisfies the equation of geodesics. This is the Euler-Lagrange equation, which is satisfied by minimizers for a wide class of integral functionals. We need this computation for motivations only, and we advise the reader who hates coordinate computations to skip it. Another possibility is to go directly to Section 6.2, and then come back to Section 5.2 and restate its material in invariant form.

Our next step is to substitute a noninvariant coordinate-wise derivative for vector fields by covariant differentiation. One also defines covariant differentiation for Finsler metrics, but its algebraic structure is much more complicated, and it is not nearly as useful and convenient as the Riemannian one. Using covariant differentiation, we rewrite the equation for geodesics and introduce the Gaussian (sectional) curvature.

Sections 6.3 and 6.4 are devoted to the geometric meaning of Gaussian curvature. Note that we suggest two alternative approaches: via *divergence of geodesics* and via *equidistant variations of curves (hypersurfaces)*. The

reader looking for geometric intuition and applications can take the results of these sections for the *definition* of sectional curvature and completely skip covariant differentiation. The latter is, however, needed for computations in actual examples, as well as to show the correctness of such "synthetic" definitions of Riemannian curvature.

6.1. Motivation: Coordinate Computations

6.1.1. Differential equations of geodesics in coordinates. Computations in this section are mainly used for motivation only, and they can be omitted or referred to later.

Variations of curves. In this chapter we will often deal with one-parameter families of curves, (also called a variation of a curve, especially if there is a designated curve in the family). A one-parameter family of curves will be usually denoted by γ_ε. Formally this is nothing but a map $(\varepsilon, t) \longrightarrow \gamma_\varepsilon(t)$ from (a region in) the coordinate (ε, t)-plane into Ω. The image of the coordinate vector field $\partial/\partial t$ under this map forms the velocity field of curves γ_ε. The image of the other coordinate field $\partial/\partial\varepsilon$ will be called the variation field. Indeed, when we vary the parameter ε keeping t fixed, $\gamma_\varepsilon(t)$ moves (away from $\gamma_0(t)$) with the velocity $(\partial/\partial\varepsilon)\gamma_\varepsilon(t)$. Notice that neither the velocity field nor the variation field is formally a vector field unless the map $\gamma_\varepsilon(t)$ is a diffeomorphism from its domain in the (ε, t)-plane into Ω: they may have "multiple values" at some points of Ω (when curves from the family intersect), and they are not defined at the points of Ω which are not in the image of the variation. Hence these "vector fields" should rather be understood as vector-valued maps from the domain of the variation (the same way as we treat the coordinate "vector fields" for a degenerate coordinate system).

In many cases we will use one-parameter families of coordinate lines of a certain (possibly degenerate) coordinate system; in this case one of the coordinate vector fields is the field of velocities, and the other one is the variation field.

Variational computation in coordinates. Let $\gamma(t) = (x(t), y(t))$, $\gamma\colon [a,b] \to \Omega$ be a smooth shortest path parameterized by arc length, $\dot\gamma(t) = d\gamma/dt = T(t)$, $\langle T, T\rangle = 1$. Here and later on differentiation with respect to t will be denoted by a dot interchangeably with $\frac{d}{dt}$. We will include γ into a family of paths with the same endpoints and make use of the fact that the length functional restricted to this family attains its minimum at γ. Choose a vector function $V(t) = (m(t), n(t))$, $V\colon [a,b] \to \mathbb{R}^2$, $V(a) = V(b) = 0$, which will be our variation field. Fix a coordinate system (x, y) and consider the family of curves $\gamma_\varepsilon = \gamma + \varepsilon V$ given by

$\gamma_\varepsilon(t) = (x(t) + \varepsilon m(t), y(t) + \varepsilon n(t))$. Of course, this is a very noninvariant construction. It is based on a particular choice of coordinates (and hence a vector structure) on Ω; but making it invariant would bring us too far away from the topic of this section. A reader familiar with variational methods on manifolds can easily convert our arguments into a coordinate-free form.

All curves γ_ε have the same endpoints, and the fact that γ is a shortest path between its endpoints implies that $(d/d\varepsilon)_{|\varepsilon=0} L(\varepsilon) = 0$, where $L(\varepsilon)$ is the length of γ_ε. Thus we have

$$0 = \frac{d}{d\varepsilon} L(\varepsilon)_{|_{\varepsilon=0}} = \frac{d}{d\varepsilon} \int_a^b \sqrt{\left\langle T + \varepsilon \dot{V}(t), T + \varepsilon \dot{V}(t) dt(t) \right\rangle}\, dt.$$

Using $\left\langle T + \varepsilon \dot{V}(t),\, T + \varepsilon \dot{V}(t) \right\rangle = 1 + 2\varepsilon \left\langle T,\, \dot{V}(t) \right\rangle + \varepsilon^2 \left\langle \dot{V}(t),\, \dot{V}(t) \right\rangle$ and differentiating under the symbol of integration, we get:

$$0 = \frac{d}{d\varepsilon} L(\varepsilon)_{|_{\varepsilon=0}} = \int_a^b \left\langle T(t),\, \dot{V}(t) \right\rangle dt. \tag{6.1}$$

Now we want to use integration by parts. It is very tempting to use the usual Product Rule for differentiating the Euclidean scalar product:

$$\frac{d}{dt} \langle T(t),\, V(t) \rangle_{\text{Eucl}} = \left\langle \dot{T}(t),\, V(t) \right\rangle_{\text{Eucl}} + \left\langle T(t),\, \dot{V}(t) \right\rangle_{\text{Eucl}},$$

but for our "scalar product" $\langle , \rangle$ this formula is just **incorrect**! Recall that we use $\langle , \rangle$ to denote a bilinear form $Q_p(\cdot, \cdot)$ on (tangent) vectors. The coefficients of this form depend on p, and hence one has to use the Chain Rule to differentiate it. There is an invariant and very convenient way of doing this by introducing *covariant derivatives* (see Section 6.2). As a matter of fact, already the expression $\dot{T}(t)$ itself depends on the choice of a coordinate system! Covariant derivatives are designed to substitute this "noninvariant" coordinate-wise derivative by an "invariant" (coordinate-independent) operation.

We will proceed here with a coordinate computation now. A geometrical meaning of what is going on here will become clear in Section 6.2.

Recall that

$$\begin{aligned} \langle T(t), V(t) \rangle = E(x(t),\, y(t)) \frac{dx}{dt} m(t) + F(x(t),\, y(t)) \frac{dx}{dt} n(t) \\ + F(x(t),\, y(t)) \frac{dy}{dt} m(t) + G(x(t),\, y(t)) \frac{dy}{dt} n(t). \end{aligned}$$

Differentiating this expression with respect to t (and omitting arguments to avoid too cumbersome formulas), we get

$$\frac{d}{dt} \langle T,\, V \rangle = \left\langle \dot{T}, V \right\rangle + \left\langle T,\, \dot{V} \right\rangle + mA + nB,$$

where

$$A = \dot{x}^2\frac{\partial E}{\partial x} + \dot{x}\dot{y}\frac{\partial E}{\partial y} + \dot{y}^2\frac{\partial F}{\partial y} + \dot{x}\dot{y}\frac{\partial F}{\partial x},$$
$$B = \dot{x}^2\frac{\partial F}{\partial x} + \dot{x}\dot{y}\frac{\partial F}{\partial y} + \dot{y}^2\frac{\partial G}{\partial y} + \dot{x}\dot{y}\frac{\partial G}{\partial x}.$$

Here the expressions A and B came from differentiating the metric coefficients $E(x(t), y(t))$, $F(x(t), y(t))$, $G(x(t), y(t))$ using the Chain Rule. We suggest that a reader make all intermediate calculations.

Thus (6.1) can be rewritten as

$$(6.2)\quad 0 = \int_a^b \left\langle T(t), \dot{V}(t)\right\rangle dt = \langle T, V\rangle\,|_a^b - \int_a^b (\left\langle \dot{T}, V\right\rangle + mA + nB)\,dt =$$
$$= \langle T, V\rangle\,|_a^b - \int_a^b \left(m(A + E\frac{d^2x}{dt^2} + \frac{d^2y}{dt^2}F) + n(B + F\frac{d^2x}{dt^2} + G\frac{d^2y}{dt^2})\right) dt.$$

Note that $\langle T, V\rangle\,|_a^b = 0$ since we supposed that all curves γ_ε have the same endpoints. Recall that this identity is true for *all* choices of functions m, n, with the only restriction that they vanish at a and b. Using a standard analytical argument one shows that this implies that both expressions

$$A + E\frac{d^2x}{dt^2} + F\frac{d^2y}{dt^2} \qquad \text{and} \qquad B + F\frac{d^2x}{dt^2} + G\frac{d^2y}{dt^2}$$

must be identically zero! Indeed, if for instance the first expression was nonzero at some point t_0, one chooses $n = 0$ and a nonnegative function m such that $m(t_0) = 1$ and m vanishes outside a small neighborhood of t_0. It is easy to see that (6.2) is not satisfied for this choice of m, n.

Hence we see that our path γ satisfies just the system of two nonlinear second-order differential equations (5.10) and (5.11) that we used to define geodesics:

$$E\frac{d^2x}{dt^2} + F\frac{d^2y}{dt^2} = -\left((\frac{dx}{dt})^2\frac{\partial E}{\partial x} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial E}{\partial y} + (\frac{dy}{dt})^2\frac{\partial F}{\partial y} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial F}{\partial x}\right),$$
$$F\frac{d^2x}{dt^2} + G\frac{d^2y}{dt^2} = -\left((\frac{dx}{dt})^2\frac{\partial F}{\partial x} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial F}{\partial y} + (\frac{dy}{dt})^2\frac{\partial G}{\partial y} + \frac{dx}{dt}\frac{dy}{dt}\frac{\partial G}{\partial x}\right).$$

Exercise 6.1.1. Repeat the same computation for a conformal metric in Ω (given by a function $\lambda(p)$) for a shortest curve γ parameterized by *Euclidean arc length*. Show that the Euclidean curvature $k(t)$ of γ satisfies the relation

$$k(t) = \left\langle T(t), \frac{\nabla\lambda(\gamma(t))}{\lambda(\gamma(t))}\right\rangle.$$

This equation says that a shortest path must bend towards the direction where λ *increases*. Although this may sound like nonsense at first glance (for a shortest path should try to go through small values of λ), the following consideration relieves this feeling. Imagine that you walk around a swamp

(with λ being very big at the center of the swamp). Suppose that the swamp is on your left. Notice that you also keep turning to your left, that is *towards* the swamp.

6.2. Covariant Derivative

The goal of this section is to substitute noninvariant component-wise differentiation (which caused us a lot of trouble due to the absence of a product rule) by an invariant operation of covariant differentiation.

When we were deducing a differential equation for geodesics, our computations became rather messy once we had to differentiate a scalar product. Moreover, we noticed that we were doing coordinate computations, and certain objects had no geometrical meaning.

Our main difficulty was that we needed to differentiate *vector* functions that take values in the tangent space. Component-wise differentiation of such functions is not invariant under coordinate changes.

Let us describe this situation in more detail. Let $\gamma(t)$ be a curve starting from $p \in \Omega$ with velocity vector T, that is, $\gamma(0) = p$, $\dot\gamma(0) = T$. Here and later on in this section we always mean that the derivatives are taken at p, that is, for $t = 0$. In order to simplify notation we do not indicate this explicitly in our formulas.

If $f\colon \Omega \to \mathbb{R}$ is a function, we differentiate it by

$$Tf = \frac{df}{dt} = \lim_{t\to 0} \frac{f(\gamma(t)) - f(\gamma(0))}{t}. \tag{6.3}$$

Now consider a vector function along γ, that is, a function $V\colon \mathbb{R} \to T\Omega$ such that $V(t) \in T_{\gamma(t)}\Omega$. The most important example of such a vector function is a restriction $V(t) = W(\gamma(t))$ of a vector field $W\colon \Omega \to T\Omega$. If we want to differentiate V, we cannot use formula (6.3): we cannot subtract tangent vectors at different points, since the result would change if we choose a different coordinate system. Of course, if we fix a coordinate system, $V(t)$ can be differentiated component-wise, as we did in Subsection 6.1.1. However, this derivative has no geometric meaning, as again it will depend on our choice of coordinates. The objective of this section is to construct an "invariant" way of differentiating vector functions. The next section contains an axiomatic definition of covariant differentiation. It is followed by two sections with motivations; these sections are optional.

6.2.1. Axiomatic definition of covariant differentiation. Let us denote the covariant differentiation that we want to construct by $\frac{D}{dt}$. Let us make a list of its desirable properties and take them as axioms. These properties are counterparts of familiar properties of the derivative for vector

functions in $\mathbb{R}^n$. We will see that these properties uniquely determine the operation of covariant differentiation.

(1). *Linearity:* For any two real constants a, b,

$$\frac{D}{dt}(aV + bW) = a\frac{D}{dt}V + b\frac{D}{dt}V.$$

(2). *Product rule:* For a real-valued function $f(t)$,

$$\frac{D}{dt}(fV) = f(p)\frac{D}{dt}V + \frac{df}{dt}V.$$

(3). *Differentiation of restrictions of vector fields:* If V is a restriction of a vector field W: $V(t) = W(\gamma(t))$, we want the value of $\frac{D}{dt}V$ to be the same for all curves γ with the same tangent vector $T = \gamma'(t_0)$ at p. This means that the following operation of (directional) covariant differentiation is well-defined: given a vector field $W: \Omega \to T\Omega$ and a vector $T \in T_p\Omega$, we set

$$\nabla_T W = \frac{D}{dt}W(\gamma(t)),$$

where γ is a curve emanating from p with velocity T.

Formally we can apply ∇_T to W only if W is a vector field defined in a neighborhood of p. However, by (3) the result $\nabla_T W$ is determined by the restriction of W to any curve with velocity T at p. The notation $\nabla_T W$ and $\frac{DW}{dt} = D/dt$ will be used interchangeably once W is defined along a curve emanating from p with velocity T. It is obvious that the operations ∇ and D/dt uniquely determine each other.

(4). *Linearity of covariant directional derivative:* For constants a, b,

$$\nabla_{aT+bS}W = a\nabla_T W + b\nabla_S W.$$

(5). *Product rule for scalar multiplication:*

$$\frac{d}{dt}\langle V, W\rangle = \left\langle \frac{DV}{dt}, W\right\rangle + \left\langle V, \frac{DW}{dt}\right\rangle.$$

Unfortunately, the covariant derivative is not uniquely defined by this list of five very natural properties. We will have to add one more property, which is a counterpart of the fact that partial derivatives commute. We will precede it by a short digression.

Commuting vector fields. Even if two vector fields form a basis at every point, they usually cannot be represented as coordinate fields of some coordinate system (even locally). Coordinate vector fields have to satisfy an additional relation: their Lie bracket must be identically zero. We will not even define the Lie bracket here; instead, when considering two vector fields simultaneously, we will restrict ourselves to a certain class of *commuting*

vector fields. Two vector fields $V, W: \Omega \to T\Omega$ are said to be commuting if they can be represented as the images of two coordinate fields under a smooth map. More precisely, we require that there exists a smooth map $\varphi: \Omega' \subset \mathbb{R}^2 \to \Omega$ such that, for every $q \in \Omega'$, $V(\varphi(q)) = X(q) = d_q\varphi(\frac{\partial}{\partial x})$ and $W(\varphi(q)) = Y(q) = d_q\varphi(\frac{\partial}{\partial y})$. Loosely speaking, a pair of commuting vector fields is a pair of "coordinate vector fields in a coordinate system with possible degenerations".

To justify the term *commuting vector fields* we suggest the following easy exercise:

Exercise 6.2.1. For two commuting vector fields V, W, and a function f, show that $VWf = WVf$ (that is, they commute as differential operators).

The converse result (stating that if $VWf = WVf$ for every smooth function f, then V and W commute) is also correct (it follows from the Frobenius Theorem). Thus this property can be used as an alternative definition of commuting vector fields.

Symmetry of covariant derivative. Now we can formulate the last axiom of covariant derivative:

(6). *Symmetry:* $\nabla_V W = \nabla_W V$ for every two *commuting* vector fields V, W.

To see why this axiom actually generalizes the commutativity property of partial derivatives, we suggest the following two exercises:

Exercise 6.2.2. Two vector fields $V = (v_1(x, y), v_2(x, y))$ and $W = (w_1(x, y), w_2(x, y))$ on $\mathbb{R}^2$ commute if and only if their components v_1, v_2, w_1, w_2 satisfy certain differential relations. Find these relations and prove this.

Hint: You may use the previous exercise applied to coordinate functions.

Exercise 6.2.3. The component-wise derivative (used in vector calculus) $D_V W$ of a vector field $W = (w_1(x, y), w_2(x, y))$ on $\mathbb{R}^2$ along a vector $V = (v_1, v_2)$ is defined as

$$D_V W = (v_1 \frac{\partial w_1}{\partial x} + v_2 \frac{\partial w_1}{\partial y}, v_1 \frac{\partial w_2}{\partial x} + v_2 \frac{\partial w_2}{\partial y}).$$

Show that for commuting vector fields $D_V W = D_W V$.

Levi-Civita Lemma. The following simple Levi-Civita Lemma is considered as the main theorem in the foundations of Riemannian geometry:

Lemma 6.2.4. *The operations $\nabla_V W$ and DV/dt are uniquely defined by properties (1)–(6).*

A detailed proof of this lemma can be found in numerous textbooks, and we will limit ourselves to a sketch.

Introduce a coordinate system (x, y). For two vector fields $V = v_1 X + v_2 Y$, $W = w_1 X + w_2 Y$, where as usual X, Y are coordinate vector fields, properties (1), (2) and (4) imply that

$$\nabla_V W = (v_1 \frac{\partial w_1}{\partial x} + v_2 \frac{\partial w_1}{\partial y})X + (v_1 \frac{\partial w_1}{\partial x} + v_2 \frac{\partial w_2}{\partial y})Y \\ + v_1 w_1 \nabla_X X + v_1 w_2 \nabla_X Y + v_2 w_1 \nabla_Y X + v_2 w_2 \nabla_Y Y.$$

Hence ∇ is uniquely determined if we know three vector fields: $\nabla_X X$, $\nabla_X Y = \nabla_Y X$ (since X and Y commute!), and $\nabla_Y Y$. Thus it is enough to determine six real-valued functions:

$$\Gamma^1_{11} = \langle X, \nabla_X X\rangle, \ \Gamma^2_{11} = \langle Y, \nabla_X X\rangle, \ \Gamma^1_{21} = \Gamma^1_{12} = \langle X, \nabla_X Y\rangle,$$

$$\Gamma^2_{21} = \Gamma^2_{12} = \langle X, \nabla_X Y\rangle, \ \Gamma^1_{22} = \langle X, \nabla_Y Y\rangle, \ \Gamma^2_{11} = \langle Y, \nabla_Y Y\rangle.$$

To find these functions (in terms of E, F, G and their partial derivatives) one computes partial derivatives of the functions $E = \langle X, X\rangle$, $F = \langle X, Y\rangle$, and $G = \langle Y, Y\rangle$ using (5), thus obtaining a nondegenerate system of linear equations for Γ^k_{ij}. This proves the uniqueness part. Now it is easy to check that ∇, determined by this choice of Γ^k_{ij}, indeed satisfies axioms (1)–(6). The details are left to the reader.

The following exercise involves rather long computations:

Exercise 6.2.5. Using formulas (5.8) and (5.9), repeat the argument to find explicit coordinate formulas for the spheres and hyperbolic planes. Using (5.7), convert the formula for the component-wise Euclidean derivative (see Exercise 6.2.3) to polar coordinates.

6.2.2. Analytical motivation. Let us look again at the problem that we ran into in Subsection 6.1.1. The problem was that we could not use the product rule to differentiate the Riemannian scalar product of two vector fields. Let us fix coordinates (x, y) and consider two vector fields $V(t) = (v_1(t), v_2(t))$, $W(t) = (w_1(t), w_2(t))$. Then, applying the chain rule as in Subsection 6.1.1, we have:

$$\frac{d}{dt}\langle V, W\rangle = \left\langle \dot V, W\right\rangle + \left\langle V, \dot W\right\rangle + 2S(V, W),$$

where

$$2S(V, W) = \frac{dE}{dt}v_1 w_1 + \frac{dF}{dt}v_1 w_2 + \frac{dF}{dt}v_2 w_1 + \frac{dG}{dt}v_2 w_2,$$

and $\dot V$, $\dot W$ are the component-wise derivatives $V' = (\dot v_1(t), \dot v_2(t))$, $W' = (\dot w_1(t), \dot w_2(t))$.

We notice that this formula differs from the product rule by the term $2S(V, W)$, which *does not involve derivatives of* V *and* W: S is just a symmetric bi-linear form. Hence it is tempting to make a correction by adding to $\dot{V} = dV/dt$ and $\dot{W} = dW/dt$ certain linear expressions in such a way that the product rule would hold. It is natural to "send one-half of $2S$ to $\dot{V}$ and one-half to $\dot{W}$", thus making our formula as symmetric as possible (this is the reason why we put a coefficient 2 in front of S). Thus we define covariant differentiation by the following formula:

$$(6.4) \qquad \frac{DV}{dt} = \left(\dot{v}_1 + \frac{1}{2}\left(\frac{dE}{dt}v_1 + \frac{dF}{dt}v_2\right),\ \dot{v}_2 + \frac{1}{2}\left(\frac{dF}{dt}v_1 + \frac{dG}{dt}v_2\right)\right).$$

Then of course $\left\langle \frac{DV}{dt}, W \right\rangle = \left\langle \dot{V}, W \right\rangle + S(V, W)$, and hence we get a nice product rule:

$$\frac{d}{dt}\langle V, W\rangle = \left\langle \frac{D}{dt}V, W \right\rangle + \left\langle V, \frac{D}{dt}W \right\rangle.$$

It is not at all clear if such a definition of DV/dt has any geometric meaning (that is, whether it is independent of the choice of a coordinate system). However, one can check that, if defined this way, DV/dt satisfies the properties (1)–(6) from the previous section (do this as an exercise!) Hence, by the uniqueness part of the Levi-Civita Lemma (6.2.4), DV/dt is indeed coordinate-independent.

6.2.3. Another motivation: Covariant differentiation on embedded surfaces. Let us consider our favorite example of an embedded surface. Thus we have an embedding $r\colon \Omega \to \mathbb{R}^3$, and $\langle , \rangle$ is given by $\langle V, W\rangle = \langle dr(V), dr(V)\rangle$, where the scalar product in the right-hand side of the formula is the usual scalar multiplication in $\mathbb{R}^3$. One may prefer to think of Ω as a subset in $\mathbb{R}^3$ (i.e., replace Ω by $r(\Omega)$), and regard r as a coordinate system. Then tangent vectors to Ω can be viewed as just vectors in the ambient space, and the Riemannian scalar product is nothing but the restriction of the Euclidean scalar product to vectors tangent to Ω.

Let us repeat the variational argument from Subsection 6.1.1 in this set-up. All scalar products here mean just the Euclidean scalar multiplication in $\mathbb{R}^3$. Recall that we begin with a smooth shortest path $\gamma\colon [a, b] \to \Omega$ parameterized by arc length, $d\gamma/dt = T(t)$, $\langle dr(T), dr(T)\rangle = 1$. Of course, $dr(T)$ is nothing but the velocity vector of our curve $r(\gamma)$ in $\mathbb{R}^3$. We choose a vector function $V(t) = (m(t), n(t))$, $V\colon [a, b] \to R^2$, $V(a) = V(b) = 0$, and consider a family of curves $\gamma_\varepsilon = \gamma + \varepsilon V$ given by

$$\gamma_\varepsilon(t) = (x(t) + \varepsilon m(t),\ y(t) + \varepsilon n(t)).$$

Since the length of γ_ε assumes its minimum at $\varepsilon = 0$ we have $\frac{d}{d\varepsilon} L(\varepsilon)_{|\varepsilon=0}=0$, where $L(\varepsilon)$ is the length of γ_ε. Thus we have

$$\begin{aligned} 0 &= \frac{d}{d\varepsilon} L(\varepsilon)_{|\varepsilon=0} \\ &= \frac{d}{d\varepsilon} \int_a^b \sqrt{\left\langle dr(T) + \varepsilon \frac{d}{dt}(dr(V(t))),\, dr(T) + \varepsilon \frac{d}{dt}(dr(V(t)))\right\rangle}\, dt. \end{aligned}$$

Differentiating under the symbol of integration and substituting $\varepsilon = 0$, we get:

$$\int_a^b \left\langle dr(T(t)),\, \frac{d}{dt} dr((V(t))\right\rangle dt = 0. \tag{6.5}$$

Now we *can use integration by parts*, for we *can apply* the product rule to the Euclidean scalar product! We get

$$0 = \frac{d}{d\varepsilon} L(\varepsilon)_{|\varepsilon=0} = \langle dr(T),\, dr(V)\rangle\,|_a^b - \int_a^b \frac{d}{dt}\,\langle dr(T(t)),\, dr(V(t))\rangle\; dt.$$

Notice that not every vector in $\mathbb{R}^3$ can be represented as $dr(V(t))$ for some V: $dr(V(t))$ has to be tangent to our surface. Now arguing as in Subsection 6.1.1 one concludes that $\frac{d}{dt} dr(T)$ has to be orthogonal to all vectors tangent to the surface at $r(\gamma(t))$. In other words, the acceleration of a shortest path parameterized by arc length remains orthogonal to the surface at all times. If thinking of a geodesic as a free motion of a particle confined to the surfaces, one can recognize a well-known mechanical principle that the centrifugal acceleration is orthogonal to the surface.

The only disadvantage of this argument is that it does not have much meaning in terms of intrinsic geometry of the surface. Imagine two-dimensional creatures living in the surface. Vectors in $\mathbb{R}^3$ other than those tangent to the surface have no meaning in their physical world. There were two objects that stuck out of two dimensions in our argument, namely, $\frac{d}{dt} dr(T(t))$ and $\frac{d}{dt} dr(V(t))$. Both of them are of the same nature: they came from differentiating a tangent vector field, and this derivative might have a nontrivial component orthogonal to the surface. Notice that both vectors come into our formulas multiplied by a vector tangent to the surface, and hence they can be replaced by their orthogonal projections to the plane tangent to the surface at $r(\gamma(t))$. This suggests the following strategy of defining an "intrinsic differentiation" of tangent vector fields: begin with the usual derivative in the ambient space, and then get rid of its component

orthogonal to the surfaces by projecting it to the tangent plane:

$$\frac{D}{dt}V(t) = dr^{-1}\mathrm{Proj}_{\,dr(T_{\gamma(t)}\Omega)}\frac{d}{dt}dr(V(t)).$$

If one wants to forget about the embedding r and think of Ω as a subset of $\mathbb{R}^3$, this formula takes a simpler form:

$$\frac{D}{dt}V(t) = \mathrm{Proj}_{\,T_{\gamma(t)}\Omega}\frac{dV(t)}{dt}.$$

It is not clear *a priori* that, if defined this way, $\frac{D}{dt}$ would not change if we consider another *isometric* embedding Ω into $\mathbb{R}^3$, that is, another embedding that induces the same Riemannian metric on Ω. To prove that $\frac{D}{dt}$ depends only on the intrinsic geometry (Riemannian metric) of Ω, one can easily check that $\frac{D}{dt}$ satisfies axioms (1)–(6) of covariant derivative, and then apply Lemma 6.2.4 (do this as an exercise!).

Second fundamental form. As a concluding remark, let us notice that the normal component of the coordinate-wise derivative is

$$\left\langle \frac{d}{dt}V, W\right\rangle - \left\langle \frac{D}{dt}V, W\right\rangle = S(V, W). \tag{6.6}$$

Comparing this with (6.4), one concludes that the normal component is a symmetric bilinear form on $T_{\gamma(0)}\Omega$; its value depends only on the values of $V(0)$ and $W(0)$. This form is called the *second fundamental form* of the surface, and its eigenvectors and eigenvalues are called the principal directions and principal curvatures. While the intrinsic geometry of a surface is determined by $\langle , \rangle$ (which is often called the *first fundamental form*), it is a well-known theorem in the differential geometry of surfaces that S and $\langle , \rangle$ together determine the shape of a surface in $\mathbb{R}^3$ (up to rigid motions).

6.2.4. Parallel transport. Though our exposition will not rely on the notion of a parallel transport of vectors, we find it relevant to give its definition and list its main properties.

When we were introducing the concept of a tangent vector, we emphasized that there is no way to define the notion of "the same direction at different points" in a consistent way. However, if we consider a region Ω with a Riemannian metric, a choice of a path γ between two points $p = \gamma(a)$ and $q = \gamma(b)$ induces a natural correspondence between $T_p\Omega$ and $T_q\Omega$. It is enough to assume that γ is piecewise smooth.

We say that a vector field along γ is *parallel* if it satisfies the following first-order linear differential equation: $DV(t)/dt = 0$. Every initial value $V(a) = W$ uniquely determines a parallel vector field V, and thus one can define a map I_γ by $I_\gamma(W) = V(b)$, $I_\gamma\colon T_p\Omega \to T_q\Omega$. This map is called

the *parallel transport along* γ. It is obvious that I_γ is a linear map since solutions of a linear differential equations depend linearly on the initial data. Moreover, I_γ is a linear isometry between $(T_p\Omega, \langle,\rangle_p)$ and $(T_q\Omega, \langle,\rangle_q)$. This follows from the fact that the scalar product of two parallel vector fields is a constant function:

$$\frac{d}{dt}\langle V(t), W(t)\rangle = \left\langle \frac{D}{dt}V(t), W(t)\right\rangle + \left\langle V(t), \frac{D}{dt}W(t)\right\rangle = 0.$$

Let us note that the Gaussian curvature (which will be introduced in the next section) is a certain measure of the dependence of I_γ on γ. Using the parallel transport, one can compute the covariant derivative in the same way as one computes the derivative of a vector-valued function to $\mathbb{R}^3$ in vector calculus:

$$\frac{D}{dt}V = \frac{d}{dt}\left(I^{-1}_{\gamma[0,t]}(V(\gamma(t)))\right).$$

Note that this is a direct analog of the formula (6.3) for differentiating scalar functions.

6.3. Geodesic and Gaussian Curvatures

6.3.1. Invariant equation of geodesics. In order to rewrite the differential equation of geodesics using the covariant derivative, let us repeat the computation from Subsection 6.1.1. Recall that we have a family of curves γ_ε with a variation vector field $V = V_\varepsilon(t) = \frac{\partial}{\partial\varepsilon}\gamma_\varepsilon(t)$, and $\gamma = \gamma_0$ is a shortest path parameterized by arc length. As usual, $T = T_\varepsilon(t) = \frac{\partial}{\partial t}\gamma_\varepsilon(t)$ is the velocity field for the family of curves γ_ε; here we do not suppose so far that curves γ_ε have the same endpoints. We have

$$\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = \int_a^b \frac{d}{d\varepsilon}_{|\varepsilon=0}\sqrt{\langle T, T\rangle}\, dt$$

(because $\langle T_0(t), T_0(t)\rangle = 1$)

$$= \int_a^b \left\langle \frac{D}{d\varepsilon}_{|\varepsilon=0} T, T\right\rangle dt = \int_a^b \langle \nabla_V T(t), T(t)\rangle\, dt.$$

Since V and T commute, by axiom (6) of the covariant derivative, we have $\nabla_V T = \nabla_T V$, and hence

$$\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = \int_a^b \left\langle T(t), \frac{D}{dt}V(t)\right\rangle dt = 0.$$

Using $\frac{d}{dt}\langle V, T\rangle = \langle \frac{D}{dt}T, V\rangle + \langle T, \frac{D}{dt}V\rangle$, we get

$$\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = \langle V, T\rangle\,|_a^b - \int_a^b \left\langle \frac{D}{dt}T(t), V(t)\right\rangle dt. \tag{6.7}$$

Now suppose that curves γ_ε have the same endpoints. Then $\langle V, T\rangle\,|_a^b = 0$. In this case $\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = 0$ for any choice of V (since γ_0 is a shortest path). Arguing as in Subsection 6.1.1, we obtain a very simple differential equation:

$$\frac{D}{dt}T(t) = 0, \text{ or in alternative notation } \nabla_T T = 0. \tag{6.8}$$

Definition 6.3.1. A path γ is called a *geodesic* if its velocity vector field T satisfies the equation (6.8).

This is very similar to the Euclidean case: a straight line is a curve traced by a motion with constant velocity. Here for a velocity to be "constant" means that its *covariant* derivative is zero.

Owing to invariant notations, equation (6.8) makes sense in any dimension.

Note that every geodesic is parameterized proportionally to arc-length (because $\frac{d}{dt}\langle T, T\rangle = 2\left\langle \frac{D}{dt}T, T\right\rangle = 0$).

Exercise 6.3.2. Prove that in dimension two Definitions 5.2.1 and 6.3.1 are equivalent.

Exercise 6.3.3. Re-prove the results of subsections 5.2.2 and 5.2.1 using the new definition of geodesics.

6.3.2. First variation of length. Geodesic curvature. Pay attention that the right side of expression (6.7) depends only on a vector field V but does not depend on a family γ_ε directly. Expression (6.7) is often called the *formula of the first variation of length*, for it describes how the length of a curve changes under a variation produced by "pulling the curve by a vector field V". Recall that it is valid only for a curve parameterized by arc length. Note that (6.7) for a Riemannian manifold has the same form as for the Euclidean plane. This reflects the fact that Riemannian manifolds are "Euclidean in the first order".

This formula takes the following simple form if γ is a geodesic:

$$\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = \langle V, T\rangle\,|_a^b. \tag{6.9}$$

This equation carries particularly nice information if we restrict ourselves to a neighborhood such that there is a unique shortest path between any two points in the neighborhood. Let $\sigma(\tau)$ be a smooth path, and denote its velocity by $V = \frac{d}{d\tau}\sigma$. Let $\rho(\tau) = d(p, \sigma(\tau))$ be the distance from p to a "moving point" $\sigma(\tau)$. Applying formula (6.7) to the family of shortest paths γ_τ connecting p and $\sigma(\tau)$ (and noticing that the corresponding variation vector field vanishes at p, for this endpoint of the geodesics γ_τ does not

move), we obtain

$$\frac{d\rho}{d\tau} = \langle V, T \rangle ,$$

where T is the tangent vector of γ_τ at its end $\sigma(\tau)$. In other words, the derivative of the distance from p to a point moving at unit speed is equal to the negative cosine of the angle between the velocity vector of the moving point and the direction from the moving point towards p (since the latter is just $-T$).

Exercise 6.3.4. Let $\sigma(t)$ be a smooth closed curve, and let $\sigma(t_0)$ be the point of σ closest to p. Show that every shortest path connecting p and $\sigma(t_0)$ is orthogonal to (the velocity of) σ at t_0.

Let us now look at the case when a curve γ is not necessarily a geodesic. Notice that the vector $\nabla_T T = \frac{D}{dt}T(t)$ is always orthogonal to $T(t)$. Indeed,

$$0 = \frac{d}{dt}\langle T(t),\, T(t)\rangle = 2\left\langle T(t),\, \frac{D}{dt}T(t)\right\rangle .$$

The number $|\frac{D}{dt}T(t)| = k_g(t)$ is called the *geodesic curvature* of γ; in dimension 2 one can assign it a sign $+$ or $-$ by choosing an orientation. In higher dimensions one often calls $S(t) = \frac{D}{dt}T(t)$ the *curvature vector*. As a matter of fact, this is nothing but the value of the second fundamental form of our curve (regarded as a one-dimensional submanifold) applied to T: compare this definition with formula (6.6). This give a new geometric insight into the meaning of the second fundamental form: it describes how the metric of a submanifold changes under variations of the submanifold. For a variation with fixed endpoints (more generally, if the variation field is orthogonal to the curve at it ends), the double substitution term $\langle V,\, T\rangle\,|_a^b$ disappears, and (6.7) takes the following form:

$$\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = \int_a^b \langle -S(t),\, V(t)\rangle\; dt. \tag{6.10}$$

Exercise 6.3.5. 1. Show that for a general parameterization of γ (not necessarily by arc length)

$$S(t) = \frac{(\nabla_T T)_N}{\langle T, T\rangle},$$

where $(\nabla_T T)_N$ is the component of the vector $\nabla_T T$ orthogonal to T.

2. If the variation field V is orthogonal to T, use a change of variables in (6.7) to show that the first variation of length is

$$\frac{d}{d\varepsilon}L(\varepsilon)_{|\varepsilon=0} = \int_a^b \left\langle \frac{\nabla_T T}{\sqrt{\langle T, T\rangle}}, V(t)\right\rangle dt.$$

Exercise 6.3.6. Carry out these computations for curves in $\mathbb{R}^2$ and $\mathbb{R}^3$ with the Euclidean metric. Compare S with a well-known notion of curvature in differential geometry. Study the length of equidistants of a closed planar curve by shrinking it towards its inward normal.

6.3.3. Curvature tensor. Choose a coordinate system (x, y), and let X and Y be (as usual) the coordinate fields. (The reader who follows higher-dimensional generalizations can take two first coordinate fields of a coordinate system $(x, y, \cdots,).$) Then of course X and Y commute as differential operators on functions: $XYf = YXf$ for every smooth function f. However, unlike the Euclidean case, in general the operators $\nabla_X = \frac{D}{dx}$ and $\nabla_Y = \frac{D}{dy}$ do not commute: for a vector field Z, $\nabla_X \nabla_Y Z \neq \nabla_Y \nabla_X Z$.

The difference $R(X,Y)Z = \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z$ is called the *curvature operator*. A real-valued expression $\langle R(X,Y)Z, W\rangle$, which depends on four vector arguments X, Y, Z, W, is called the *curvature tensor*. The definition of R suggests that it depends on four vector fields (and the first two of them are supposed to be coordinate vector fields for some coordinate system). It is clear that $R(X,Y)Z$ is a linear function in each of its three arguments. The following exercises shows that $R(X,Y)Z$ at a point p depends *only on the values of X, Y, and Z at p!* The proofs of the exercises consist of a straightforward computation based on axioms of the covariant derivative:

Exercise 6.3.7. Show that, for a smooth function f:
(1) $\nabla_X \nabla_Y (fZ) - \nabla_Y \nabla_X (fZ) = f \cdot (\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z)$,
(2) $\nabla_X \nabla_{(fY)} Z - \nabla_{(fY)} \nabla_X = f \cdot (\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z)$.

Exercise 6.3.8. Show that:
(1) $R(X,Y)Z = -R(Y,X)Z$; in particular, $R(X,X)Z = 0$,
(2) $\langle R(X,Y)Z, W\rangle = \langle R(Z,W)X, Y\rangle$.

Exercise 6.3.9. Combining the two previous exercises, show that $R(X,Y)Z = 0$ at p if at least one of the fields X, Y, Z vanishes at p. Conclude that $R(X,Y)Z$ at p depends **only on the values of X, Y, and Z at p!**

For two vectors $X, Y \in T_p\Omega$, we denote $K(X,Y) = \langle R(X,Y)X, Y\rangle$. Let us choose two vectors $X, Y \in T_p\Omega$ such that the area of the parallelogram spanned by them is 1 (where the area is taken with respect to $\langle , \rangle_p$). For instance, we can choose two orthogonal unit vectors. The quantity $K(p) = K(X,Y) = \langle R(X,Y)X, Y\rangle$ is called the *sectional* curvature at p. In dimension two it is the same as the Gaussian curvature of a surface.

Exercise 6.3.10. Using symmetry properties of R (and in particular $R(X + \alpha Y, Y)Z = R(X,Y)Z$; see Exercise 6.3.8), show that $K(X,Y)$ depends only on the area of the parallelogram spanned by X, Y, and hence $K(p)$ is well-defined.

Exercise 6.3.11. Show that if a Riemannian metric is multiplied by a constant c (that is, its metric coefficients are multiplied by c^2), then its Gaussian curvature gets divided by c^2.

Exercise 6.3.12. Use Exercise 6.2.5 to compute the Gaussian curvature of the spheres and the hyperbolic planes.

Exercise 6.3.13. Let X and Y be two orthonormal vectors. Prove that we get $R(X,Y)Z$ if we rotate $|K_\sigma|Z$ on the angle $\pi/2$ in a direction depending on sign of K_σ. Can you generalize this exercise to dimensions greater two?

The importance of the Gaussian (sectional) curvature will be explained in the following section. We conclude this section with a remark related to the higher-dimensional case. The terms "sectional curvature" and "Gaussian curvature" are more or less interchangeable in our 2-dimensional exposition and the former is perhaps even more common. On the other hand, the term "sectional curvature" sounds silly in our two-dimensional context. The reason why this curvature is called "sectional" in higher dimensions becomes clear from the following three exercises, which are addressed to a reader who keeps track of a multi-dimensional theory parallel to our two-dimensional exposition. The three exercises consider the case when the dimension of Ω is greater than two.

Exercise 6.3.14. Show that $K(X,Y) = K(X_1,Y_1)$, provided that the parallelograms spanned by X,Y and X_1,Y_1 have equal areas and belong to the same 2-plane. Hence $K(X,Y)$ depends only on $X \wedge Y$.

Let σ be a two-dimensional plane $\sigma \subset T_p\Omega$. Define K_σ by $K_\sigma = K(X,Y)$, where X and Y are two orthonormal vectors $X,Y \in \sigma$. Equivalently, $K_\sigma = \frac{K(X,Y)}{||X\wedge Y||^2}$.

Exercise 6.3.15. Consider a two-dimensional embedded surface obtained by restricting $\exp_p$ to a small neighborhood of the origin in σ; equip it with the metric induced from Ω. One can think of this surface as a two-dimensional section of Ω formed by rotating a geodesic passing through p so that its velocity vectors sweep σ. Show that the sectional curvature $K_\sigma(p)$ is equal to the sectional (Gaussian) curvature of this two-dimensional surface at p.

Exercise 6.3.16. Use linear properties of the curvature operator to find how the sectional curvature depends on a (two-dimensional) direction σ in a metric product $S^2 \times S^2$ of two unit spheres (assume that you know that the curvature of the unit sphere is 1). Do the same for other products, such as the product of Euclidean and hyperbolic planes, and a sphere and a hyperbolic plane.

Parallel transport and curvature. Gauss–Bonnet Formula. There is the following description of the curvature using the parallel transport. Let $X, Y \in T_p\Omega$ be two vectors such that the area of the parallelogram spanned by them is 1. Consider a closed path γ, $\gamma(a) = \gamma(b) = p$. Then I_γ defines a linear isometry from $T_p\Omega$ to itself. Then

$$R(X,Y)Z = (I_\gamma(Z) - Z)\text{Area}(\gamma) + o(\text{Area}(\gamma)), \tag{6.11}$$

where Area(γ) is the (oriented) Riemannian area enclosed by γ. Notice that formula (6.11) easily follows from the definition of the curvature operator if γ is a rectangle formed by coordinate lines of a coordinate system whose coordinate vectors at p are X and Y. This formula means that, when a vector is carried along a small loop back to the same point, the vector gets rotated by an angle that is close to the product of the Gaussian curvature at the point and the area enclosed in the loop. Formula (6.11) can be easily generalized to higher dimensions (in terms of differential two-forms, or by imposing an assumption formalizing that "γ encloses a region in a two-dimensional surface tangent to X and Y"). For two-dimensional surfaces formula (6.11) implies the following Gauss–Bonnet Formula:

Theorem 6.3.17. *Let a smooth closed curve γ enclose a simply-connected region Ω. Then*

$$\int_\Omega K(p)\,dA(p) = \int_\gamma k_g(t)\,dt - 2\pi,$$

where γ is assumed to be parameterized by arc length, and $\int_\Omega K(p)\,dA(p)$ is the integral of the Gaussian curvature with respect to the Riemannian area.

If γ is a piecewise smooth curve, then one has to add $\pi - \alpha_i$ for each corner with interior angle α_i to the right-hand side of the formula.

To get a better grasp of this theorem, apply it to (1) a smooth planar curve, (2) a Euclidean triangle (3), to a greater circle in a sphere, and (4) to each of the two regions whose boundary is a small circle in a sphere.

We will not use formula (6.11) and Theorem 6.3.17, and their proofs are omitted. A proof of the Gauss–Bonnet Formula for hyperbolic and spherical triangles can be found in Section 5.3.

6.4. Geometric Meaning of Gaussian Curvature

The main property that distinguishes a Riemannian length structure from other length spaces is that locally it is "almost Euclidean". More precisely, Lemma 5.1.13 asserts that every point p admits a neighborhood with coordinate system such that the metric distortion (with respect to the auxiliary Euclidean metric induced by the coordinate system) is of a smaller order than the distance to p. We are going to refine this assertion by showing that

(appropriately chosen) normal coordinates do not distort the metric in the second order, and the third-order distortions are controlled by the Gaussian curvature.

There are two ways of looking at metric distortions in normal coordinates. Recall that one family of coordinate lines is formed by geodesics, and hence the metric along these lines is preserved. Hence we can conclude that the metric distortion arises from the difference in the behavior of the distance between near-by geodesics in the Riemannian and Euclidean spaces. We will see that the divergence of geodesics is governed by a second-order differential equation whose difference from the corresponding Euclidean equation is a curvature term. This approach is realized in the subsection "Jacobi equation". An advantage of this approach is that general definitions of curvature bounds for length spaces are given in the same terms.

Alternatively, we can recall that the other family of coordinate lines is orthogonal to geodesic coordinate lines. We will see that this is a family of equidistant curves, with a natural correspondence between the points on different equidistant curves along the geodesic coordinate lines. Thus the metric distortion can be measured by the distortion of the length of equidistant curves as we move along the geodesic family of coordinate lines. We will see that the first derivative of length of an equidistant curve is given by its geodesic curvature, and the second derivative is again governed by the Gaussian curvature of the metric. This approach is realized in the subsection "Equidistant families". An advantage of this approach is that its generalization to higher dimensions (replacing equidistant lines by hypersurfaces) is almost straightforward. Unfortunately, there is no obvious analog of this method in more general length spaces.

Our further geometric conclusions can be derived from either of the methods.

6.4.1. Jacobi equation. Let us consider a normal coordinate system (x, y) with coordinate vector fields X, Y. Recall that the x-lines are geodesics parameterized by arc length and orthogonal to the y-lines. This means that $\nabla_X X = \frac{D}{dx} X = 0$ (we use both notations for the covariant derivative to refresh notations). Let us differentiate this identity with respect to y:

$$\nabla_Y \nabla_X X = \frac{D}{dy}\frac{D}{dx} X = 0.$$

We want to change the order of differentiations. The covariant derivatives with respect to x and y do not commute, and the curvature operator has been introduced in the previous section just as their commutator. Thus

$$R(X, Y)X = \nabla_Y \nabla_X X - \nabla_X \nabla_Y X = -\nabla_X \nabla_Y X = -\nabla_X \nabla_X Y,$$

where the last identity follows from axiom (6) of covariant differentiation: $\nabla_X Y = \nabla_Y X$ for commuting (coordinate) vector fields X, Y. We will denote $\nabla_X \nabla_X = \frac{D}{dx}\frac{D}{dx}$ by $\frac{D^2}{dx^2}$.

Thus we get the following *Jacobi equation*, which is a second-order linear differential equation for Y:

$$\frac{D^2}{dx^2} Y = -R(X, Y)X. \tag{6.12}$$

Notice that Y is nothing but a variational vector field for a family of geodesics (the x-lines); it is of no importance that these lines form a nondegenerate coordinate system, and the same argument can be repeated verbatim for a general variation. Let us give the corresponding definition:

Definition 6.4.1. A vector field $Y(t)$ along a geodesic $\gamma(t)$ is called a *Jacobi vector field* if it satisfies equation (6.12).

Exercise 6.4.2. Show that every Jacobi field along γ can be represented as a variational vector field by including γ into a family of geodesics.

Hint: Since a Jacobi vector field Y satisfies a second-order linear differential equation, it is uniquely determined by its Cauchy data, that is, its value $Y(t_0)$ and its derivative $\frac{D}{dt}Y(t_0)$ at a point t_0. On the other hand, it is easy to include γ into a geodesic variation whose variation vector field Y_1 has the same value and the same derivative at t_0 as Y. This implies that $X = Y$, for Y has to satisfy the same equation (6.12), since it is a variation field of a family of geodesics.

Let us consider a Jacobi field $Y(t)$ along a geodesic $\gamma(t)$. As usual, denote the velocity vector of γ by T. Denote by $g(t)$ the length of $Y(t)$. Suppose that $g(t) \neg 0$. Then one can write $g(t)) = |Y(t)| = \langle Y(t), N(t) \rangle$, where $N = |Y|^{-1}Y$ is a unit vector field orthogonal to γ. The function g characterizes how geodesics from a family with variational vector Y diverge from γ: the distance between $\gamma(t)$ and $\gamma_\varepsilon(t)$ is approximately $\varepsilon g(t)$ (up to $o(\varepsilon)$).

Remark. Our definition of N assumes that $|Y| \neq 0$, and it can be smoothly extended to the closure of the interval where g and Y do not vanish. Hence all further conclusions in this subsection are valid *only as long as t belongs to this interval.*

The following obvious lemma will be quite handy:

Lemma 6.4.3. *Covariant derivatives of a unit vector field are orthogonal to this vector field: if W is a unit vector field, then $\langle \nabla_Z W, W \rangle = 0$, where Z is any vector field.*

Proof. By differentiating the expression $\langle W, W\rangle = 1$ we get

$$0 = Z\langle W, W\rangle = 2\langle \nabla_Z W, W\rangle .$$

□

Let us make some preliminary computations. By differentiating the expression $\langle N, T\rangle = 0$, we obtain that

$$0 = \frac{d}{dt}\langle N, T\rangle = \left\langle \frac{D}{dt}N, T\right\rangle + \left\langle \frac{D}{dt}T, N\right\rangle = \left\langle \frac{D}{dt}N, T\right\rangle,$$

since γ is a geodesics and hence by definition $\frac{D}{dt}T = 0$. Thus we have

$$\left\langle \frac{D}{dt}N, Y\right\rangle = \left\langle \frac{D}{dt}N, T\right\rangle = 0 \quad \text{and} \quad \left\langle \frac{D}{dt}N, N\right\rangle = 0, \tag{6.13}$$

where the last identity follows from Lemma 6.4.3.

We want to compute the second derivative of g. Using (6.13), we get

$$\frac{dg}{dt} = \frac{d}{dt}\langle Y, N\rangle = \left\langle \frac{D}{dt}Y, N\right\rangle .$$

One more differentiation yields

$$\frac{d^2 g}{dt^2} = \frac{d}{dt}\left\langle \frac{D}{dt}Y, N\right\rangle = \left\langle \frac{D^2}{dt^2}Y, N\right\rangle + \left\langle \frac{D}{dt}Y, \frac{D}{dt}N\right\rangle .$$

Using $\frac{D}{dt}Y = \frac{D}{dt}(gN) = \frac{dg}{dt}N + g\frac{D}{dt}(N)$, we can rewrite this identity as

$$\frac{d^2 g}{dt^2} = \frac{d}{dt}\left\langle \frac{D}{dt}Y, N\right\rangle = \left\langle \frac{D^2}{dt^2}Y, N\right\rangle + g\left\langle \frac{D}{dx}N, \frac{D}{dx}N\right\rangle .$$

Notice that, by the Jacobi equation (6.12),

$$\begin{aligned}\left\langle \frac{D^2}{dt^2}Y, N\right\rangle &= -\langle R(T,Y)T, N\rangle \\ &= -\langle R(T,gN)T, N\rangle = -g\langle R(T,N)N, T\rangle = -gK,\end{aligned}$$

where K is the Gaussian (sectional) curvature (and the last equality uses the fact that T and N are orthogonal unit vectors). Hence we finally get that the Jacobi equation implies the following differential equation for the scalar function g:

$$\frac{d^2 g}{dt^2} = -g(K + \left\langle \frac{D}{dx}N, \frac{D}{dx}N\right\rangle), \tag{6.14}$$

where $K = K(\gamma(t))$ is the Gaussian curvature along γ.

Let us notice now that actually $\frac{D}{dt}N = 0$, for $\frac{D}{dt}N$ is orthogonal to both T and N, and we are dealing with the two-dimensional case! Hence in the two-dimensional case equation (6.14) turns into

$$\frac{d^2 g}{dt^2} = -gK. \tag{6.15}$$

The reason why we postponed this observation till now and derived the more complicated equation (6.14) is explained in the following remark:

Important remark. Now we have arrived at the point where there is an essential difference between two-dimensional surfaces and the higher-dimensional case. In the two-dimensional case the direction of Y is determined by the fact that Y is orthogonal to T. Hence we were able to transform the Jacobi equation (6.12) into one scalar equation for the magnitude of Y. In higher dimensions the Jacobi equation (6.12) involves both the magnitude and the direction of Y. The higher-dimensional situation is complicated by the following phenomenon: geodesics from our family "twist around γ". It is important to notice that the equation for the length of Y that we obtained for the two-dimensional case is *incorrect* in higher dimensions! However, the main geometrical corollary, the Rauch Comparison Theorem 6.5.1, is still true. One of the proofs consists of a computation based on decomposing y with respect to a basis of *parallel vector fields along* γ, that is, such vector fields that their covariant derivatives along γ are zero. One such vector field along γ is T, and in our two-dimensional argument N just happened to be another parallel vector field along γ. The next section gives an alternative approach via equidistant variations. This approach can be easily generalized to higher dimensions, and it can be used instead of the Rauch Comparison Theorem to prove the Cartan–Alexandrov–Toponogov Comparison Theorem. Note also that many geometric corollaries can be extracted directly from the inequality $\frac{d^2 g}{dx^2} \le -gK$ implied by (6.14).

Equidistant variations. The previous section described the Gaussian curvature as a quantitative way of measuring how a family of geodesics diverges. There is an alternative approach, which avoids using the Jacobi equation. It has certain advantages for Riemannian geometry (and in particular it avoids the complication in higher dimensions that we ran into when deriving (6.15)). On the other hand, this approach can hardly be used for length spaces other than manifolds.

The idea of this approach is that, instead of studying a bunch of geodesics, we will concentrate our attention on the family of lines orthogonal to them. In a normal coordinate system this is just the family of the y-lines. The distance between two nearby geodesic lines is approximately the length of the segment of the orthogonal line enclosed between them. This suggests

that we study how the length along an orthogonal line changes within its family. A convenient way of doing so is to consider this family of orthogonal lines as a variation of a curve (in higher dimensions one studies variations of hypersurfaces).

Let γ_ε be a family of curves whose variation vector field V is a unit vector field orthogonal to the curves of the family: $\langle V, V\rangle = 1$, and $\langle V, T\rangle = 0$ (where as usual $T = \frac{d}{dt}\gamma$, and we drop the arguments t, ε to simplify notation). Such variation is said to be equidistant. The reason is explained in the following exercise.

Exercise 6.4.4. Prove that, for a fixed t_0,

$$\inf_t\{d(\gamma_\varepsilon(t_0), \gamma_0(t_0))\} = \varepsilon$$

for all sufficiently small ε. In other words, points of γ_ε lie at the same distance ε from (the image of) γ_0.

Hint: Use the fact that the lines $\sigma(\varepsilon) = \gamma_\varepsilon(t)$ are geodesics (see below), and argue as in the proof of the Gauss Lemma.

One should not think of the curves γ_ε as geodesics, but rather as lines orthogonal to a family of geodesics. Indeed, statement (ii) of Lemma 6.4.6 will imply that the orthogonal curves $\sigma(\varepsilon) = \gamma_\varepsilon(t)$ are geodesics.

Let us assume that the line $\gamma = \gamma_0$ is parameterized by arc length.

According to the formula of the first variation (6.10),

$$\frac{d}{d\varepsilon}L(\gamma_\varepsilon, a, b) = \int_a^b \langle -S, V\rangle\, dt = \int_a^b -k_g(t)\, dt,$$

where $S = \nabla_T T$, and $k_g = \langle \nabla_T T, V\rangle$ is the geodesic curvature of γ at t (and its sign is chosen with respect to the orientation prescribed by V). Hence a curve in our family shrinks (if $k_g > 0$) or expands ($k_g < 0$) at a rate proportional to its curvature.

Exercise 6.4.5. Show that for a variation with a variation field fV, where V is a unit normal to the curve and f is a scalar function,

$$\frac{d}{d\varepsilon}L(\gamma_\varepsilon, a, b) = \int_a^b -f(t)k_g(t)\, dt.$$

We will need the following computational lemma:

Lemma 6.4.6. *In notation before Exercise 6.4.4, one has:*

(i) $\langle \nabla_T V, V\rangle = \langle \nabla_V V, V\rangle = 0$;

(ii) $\nabla_V V = 0$, *(and therefore the lines orthogonal to the family* γ_ε *are geodesics);*

(iii) $\langle \nabla_T \nabla_V T, V\rangle = -\langle \nabla_T V, \nabla_T V\rangle$;

(iv) $\langle \nabla_T V, \nabla_T V \rangle = \langle \nabla_T T, \nabla_T T \rangle = k_g^2$.

Proof. (i) This is true just because V and T are unit vector fields (Lemma 6.4.3).

(ii) Since $\langle V, T \rangle = 0$, we have

$$0 = \frac{d}{d\varepsilon} \langle V, T \rangle = \langle \nabla_V V, T \rangle + \langle V, \nabla_V T \rangle$$
$$= \langle \nabla_V V, T \rangle + \langle V, \nabla_T V \rangle = \langle \nabla_V V, T \rangle + \frac{1}{2}\frac{d}{dt} \langle V, V \rangle = \langle \nabla_V V, T \rangle .$$

Together with (i) this implies that the vector $\nabla_V V$ is zero, for it is orthogonal to both V and T.

(iii) Differentiating $\langle \nabla_T V, V \rangle = 0$ (which follows from (i)), we get

$$0 = \frac{d}{dt} \langle \nabla_T V, V \rangle = \langle \nabla_T \nabla_V T, V \rangle + \langle \nabla_T V, \nabla_T V \rangle .$$

Together with $\nabla_T V = \nabla_V T$ (since T and V are commuting vector fields) this implies (iii).

(iv) Since $|k_g| = |\nabla_T T|$, the equality $\langle \nabla_T T, \nabla_T T \rangle = k_g^2$ is trivial. Differentiating $\langle V, T \rangle = 0$, we get

$$0 = \frac{d}{dt} \langle T, V \rangle = \langle \nabla_T T, V \rangle + \langle T, \nabla_T V \rangle .$$

Since V and T are unit vectors, $\nabla_T T$ is proportional to V, and $\nabla_T V$ is proportional to T, this implies that $|\nabla_T T| = |\nabla_T V|$, which proves (iv). □

Remark. Although it may seem that the argument for (ii) uses the fact that $\dim(\Omega) = 2$ in the same way as in the computation used to derive (6.15), it is not the case. We leave as an exercise to show that the integral curves of a variation field V of an equidistant variation of a hypersurface are geodesics. To do this, one just shows that $\nabla_V V$ is orthogonal to V and to all vectors tangent to the hypersurfaces. This is a reason why this approach is sometimes more convenient in higher dimensions.

Now we want to look at the second derivative of length; that is, we want to know how k_g changes as we change the parameter ε. We have

$$\frac{dk_g}{d\varepsilon} = \frac{d}{d\varepsilon} \langle \nabla_T T, V \rangle = \langle \nabla_V \nabla_T T, V \rangle + \langle \nabla_T T, \nabla_V V \rangle = \langle \nabla_V \nabla_T T, V \rangle ,$$

since $\nabla_V V = 0$ by Lemma 6.4.6, (ii).

Now we want to change the order of differentiations in $\langle \nabla_V \nabla_T T, V \rangle$, and thus the curvature operator shows up. We get

$$\frac{dk_g}{d\varepsilon} = \langle \nabla_T \nabla_V T, V \rangle + \langle R(T, V)T, V \rangle .$$

The second term is Gaussian curvature (since V, T form an orthonormal base), and the first term is $-k_g^2$ by Lemma 6.4.6, (iii) and (iv). Thus we proved the following differential equation for k_g (called a "Ricatti equation").

Lemma 6.4.7.

$$\frac{dk_g}{d\varepsilon} = -k_g^2 + K.$$

Remark. The statement of Lemma 6.4.7 can be used as an alternative definition of the Gaussian curvature.

Exercise 6.4.8. Apply Lemma 6.4.7 to the inwards equidistant variation of a Euclidean circle (remember that the choice of the sign for k_g depends on the direction of the variation). Notice that when ε is equal to the radius of the circle, $k_g = \infty$ (the circle shrinks to a point).

Exercise 6.4.9. Apply Lemma 6.4.7 to the outwards equidistant variation of a circle in a plane of constant Gaussian curvature -1. What is the limit of k_g as $\varepsilon \to \infty$? Do the same for a circle on a sphere.

The next exercise allows us to compare the geodesic curvatures of equidistant families in two regions provided that it is known that the Gaussian curvature in one region does not exceed that in the other:

Exercise 6.4.10. Let $\gamma_\varepsilon(t)$, $\sigma_\varepsilon(t)$ be two equidistant families of smooth curves in Ω, Ω_1 (with the variation vectors orthogonal to the curves). We assume that the variations are defined in the same domain in the (ε, t)-plane. Assume that the geodesic curvatures of γ_0 and σ_0 are the same at t_0, and that the Gaussian curvature in Ω and Ω_1 satisfies the inequality $K(\gamma_\varepsilon(t_0)) \geq K(\sigma_\varepsilon(t_0))$ for all ε in the domain of the variations. Prove that, for every ε from the domain of the variations, the geodesic curvature of γ_ε is less than or equal to the geodesic curvature of σ_ε at t_0.

This exercise is a counterpart of Theorem 6.5.1, and it can be used instead of it to prove the main Theorem 6.5.4. Notice that Theorem 6.5.1 has a condition that $Y(t)$ does not vanish; the statement of the theorem does not hold past the point where $Y(t)$ vanishes (this point is said to be *conjugate* to $\gamma(0)$ along γ). A counterpart of this condition in Exercise 6.4.10 is the assumption that all curves in the equidistant family are smooth; that is, k_g does not become infinite, as happens for the inwards equidistants of a circle when the equidistants collapse to a point (which is the center of the circle). A point where an equidistant of a curve acquires a singularity $k_g = \infty$ is said to be *focal*.

Exercise 6.4.11. Show that, for a planar curve $\gamma(t)$, its focal points form the curve $\gamma(t) + (1/k_g(t))\, N(t)$, where $N(t)$ is a unit normal vector field.

The number $1/k_g(t)$ is called the *curvature radius of* γ *at* $\gamma(t)$.

There is the following fruitful strategy of studying geometry (or topology) of a Riemannian manifold with certain curvature bounds using Lemma 6.4.7 (or the result of Exercise 6.4.10): one sweeps through a manifold by a family of equidistant hypersurfaces (curves in dimension two), and then compares this family with an appropriate equidistant family in a model space (usually of constant curvature).

6.4.2. Metric distortion of the exponential map. The results of this subsection can be based on either of the two previous subsections. We will use the Jacobi equation, leaving carrying out the other approach as an exercise.

Let us apply equation (6.15) to the family of the x-lines of a normal coordinate system centered at p. Recall that our metric takes a very simple form in normal coordinates: $E = \langle X, X\rangle = 1$, $F = \langle X, Y\rangle = 0$, and $G = \langle Y, Y\rangle$ is the only nontrivial coefficient. Using the notation $g = |Y| = \sqrt{G}$, we conclude that

$$\frac{\partial^2\sqrt{G}}{\partial x^2} = -K\sqrt{G}.$$

This equation describes the divergence of two nearby coordinate lines from each other.

More generally, if a geodesic $\gamma(t)$ is included into a family of geodesics $\gamma_\varepsilon(t)$, $\gamma = \gamma_0$, and the variation vector $Y(t) = \frac{\partial}{\partial\varepsilon}\gamma_\varepsilon(t)|_{\varepsilon=t=0}$ is orthogonal to $\dot\gamma(0)$, then the length $g(t)$ of this vector satisfies

$$\frac{d^2g(t)}{dt^2} = -K(t)g(t), \tag{6.16}$$

where $K(t) = K(\gamma(t))$ is the Gaussian curvature along γ. Recall that this equation is valid only as long as $g(t)$ does not vanish; vanishing of $g(t)$ geometrically means that nearby geodesics from the family meet γ (at least up to the second order).

For a normal coordinate system centered at p we have $g(0) = G(0, y) = 0$, for the coordinate vector Y vanishes at p. Notice that $dg/dt(0) = 1$. This is a more delicate statement: it follows from Lemma 5.1.13 (two x-lines emanating from p diverge in the first order at the same rate as two Euclidean rays emanating from one point and forming the same angle). Hence, by (6.16), $d^2g/dt^2(0) = 0$ and therefore $g(t) = t + o(t^2)$. Plugging this again in (6.16), we get

$$\frac{d^2g}{dt^2}(t) = -Kt + o(t^2) \quad \text{and hence} \quad \frac{d^3g}{dt^3}(0) = -K.$$

In other words, in normal coordinates

$$\sqrt{G(x,y)} = x - \frac{1}{6}Kx^3 + o(x^3). \tag{6.17}$$

This conclusion gives a very good idea of a geometric meaning of curvature: in normal coordinates the metric is Euclidean up to the second order (since $\sqrt{G(x,y)} = x$ for a Euclidean metric), and the third-order distortion is measured by K.

Let us give another reformulation of the assertion that a Riemannian metric is locally Euclidean up to a third order error. Recall that a Riemannian metric locally can be approximated by a Euclidean metric as in Lemma 5.1.13. This lemma guarantees that one can choose a (coordinate) map from a Euclidean region to a neighborhood of a given point p in Ω in such a way that its metric distortion in a ball of (a small) radius r centered at p is $o(r)$. Now we can essentially refine this result. The tangent space $T_p\Omega$ together with $\langle , \rangle_p$ is a Euclidean space, and $\exp_p$ maps a neighborhood of the origin in this space to Ω. The following lemma asserts that this map differs from an isometry by at most a third-order term (in the distance from p):

Lemma 6.4.12. *For a unit vector $V \in T_q\Omega$, we have*

$$d(q,s) - |\exp_p^{-1}(q)\exp_p^{-1}(s)| = o(\max(d(p,q), d(p,s))^3),$$

where $|\cdot|$ is the distance in $(T_\Omega, \langle , \rangle)$.

Proof. Introduce a polar coordinate system (r,ρ) in $T_p\Omega$ and choose a normal coordinate system in a neighborhood of p such that $\exp_p$ maps points in $T_p\Omega$ to points in Ω with the same coordinates: $\exp_p(r,\rho) = (x = r, y = \rho)$. Then the metric coefficients of the Euclidean metric $\langle , \rangle_p$ in $T_p\Omega$ take the form $E = 1, F = 0, G(r,\rho) = r^2$, while the metric coefficients of our Riemannian metric at the corresponding point $x = r, y = \rho$ are $E = 1, F = 0, G(r,\rho) = (r + o_\rho(r^2))^2$. Now the lemma follows from formula (5.3) for the Riemannian length of a curve in coordinates. □

Let us notice again that the Gaussian curvature measures just the third-order distortion of metric.

The construction used to prove this lemma is very useful, and it will be applied (with minor modifications) in this chapter several times! Notice also that $\exp_p$ does not distort distances from p, and hence it maps spheres centered at the origin of $T_p\Omega$ to spheres centered at p.

Surfaces of constant curvature. Let us consider an important model situation of a region of a constant Gaussian curvature $K(p) = k$ for all $p \in \Omega$. There are three cases: $k = 0$, $k > 0$ and $k < 0$, which correspond to

surfaces that are locally isometric to the Euclidean plane, a sphere, and a hyperbolic plane. Indeed, if $k = 0$, the Jacobi equation (6.16) tells us that $\frac{d^2}{dt^2}g(t) = 0$, and hence $g(t)$ is a linear function. Indeed, Euclidean geodesics are straight lines, and the distance between straight lines is a linear function (as long as it does not vanish!) If $k > 0$, one explicitly solves the Jacobi equation (6.16), obtaining its general solution $g(t) = A\sin\sqrt{k}\,t + B\cos\sqrt{k}\,t$. In particular, if $g(0) = 0$, $\dot{g}(0) = 1$, we get $g(t) = \sin\sqrt{k}\,t$. It is easy to check that if one rotates a meridian of a sphere of radius $\sqrt{k}$ around its pole at the unit angular speed, the speed of a point of this meridian lying at a distance t from the pole is $\sin\sqrt{K}\,t$; in other words, nearby meridians diverge at a speed proportional to $\sin\sqrt{k}\,t$. If $k < 0$, the general solution of equation (6.16) is $g(t) = A\sinh\sqrt{-k}\,t + B\cosh\sqrt{-k}\,t$. This formula describes the divergence of a bunch of lines in the hyperbolic plane of curvature k. Let us formalize these considerations:

Lemma 6.4.13. *Let $K(q) = k$ for all $q \in \Omega$. Let $p \in \Omega$. Then there is a neighborhood of p that is isometric to a region in the Euclidean plane ($k = 0$), the hyperbolic plane of curvature k ($k < 0$), or the sphere of radius $\sqrt{k}$ ($k > 0$)*

We consider the case $k = 0$; the two other cases are similar. The argument is very similar to the proof of Lemma 6.4.12. Choose a normal coordinate system (x, y) centered at p. Let $\gamma(t) = (t, y)$ be a coordinate line emanating from p. Denote, as usual, $g(t) = \sqrt{G(t, y)}$. As in the proof of Lemma 6.4.12, $g(0) = 0$ and $\frac{dg}{dt}(0) = 1$. Solving the Jacobi equation (6.16) subject to the initial conditions $g(0) = 0$, $g'(0) = 1$, we immediately get $g(t) = t$, and hence $G(x, y) = y^2$. Comparing these expressions for metric coefficients E, F, G with the metric coefficients of the Euclidean metric in polar coordinates (r, ρ) (formula (5.7)), we see that the map $(x, y) \to (r = x, \rho = y)$ is an isometry to a neighborhood of the origin in the Euclidean plane in polar coordinates. □

Exercise 6.4.14. Using formulas (5.8) and (5.9), repeat the same argument to show that a surface of constant curvature is locally isometric to a sphere or a hyperbolic plane.

Remark. Notice that we *proved* that the Gaussian curvature of the hyperbolic plane of curvature k is k; moreover, our argument also *implies* that the hyperbolic planes are homogeneous: notice that we constructed an isometry mapping any given point in a region of constant negative curvature to the origin of the polar coordinates used in formula (5.9)!

6.5. Comparison Theorems

6.5.1. Rauch Comparison Theorem. The following (two-dimensional case of the) Rauch Comparison Theorem (or its analog Exercise 6.4.10) plays a crucial role for our metric considerations.

Theorem 6.5.1. *Let $\gamma_1(t), \gamma_0(t)$ be geodesics in Ω_1, Ω_0 respectively. Let $Y_1(t)$, $Y_0(t)$ be Jacobi fields along γ_1 and γ_0 such that*

$$Y_1(0) = Y_0(0) = 0 \quad \textit{and} \quad |\frac{D}{dt}Y_1(0)| = |\frac{D}{dt}Y_0(0)| = 1.$$

Let us assume that Y_1 does not vanish on an interval $[0, T]$, and that the Gaussian curvature $K_1(t) = K(\gamma_1(t))$ of Ω_1 along γ_1 is less than or equal to the Gaussian curvature $K_0(t) = K(\gamma_0(t))$ of Ω_0 along γ_0, that is, $K_0(t) \leq K_1(t)$ for all $t \in [0, T]$.

Then $|Y_1(t)| \leq |Y_0(t)|$ for all $0 \leq t \leq T$.

Proof. Let $g_1(t) = |Y_1(t)|$, $g_0(t) = |Y_0(t)|$. Then $g_1(t)$ satisfies the equation

$$\frac{d^2 g_1}{dt^2}(t) = -K_1(t) g_1(t) \quad \text{subject to} \quad g_1(0) = 0,\ \dot{g}_1(0) = 1,$$

and $g_0(t)$ satisfies the similar equation.

We want to prove that $g_1(t) \leq g_0(t)$ for all $0 \leq t \leq T$.

Consider a family of functions $g_s(t)$, $1 > s \geq 0$, determined by the equation

$$\frac{d^2 g_s}{dt^2}(t) = -(K_0(t) - s) g_s(t) \quad \text{subject to} \quad g_s(0) = 0,\ \dot{g}_s(0) = 1 - s.$$

Notice that for $s = 0$ the notation is consistent. We are going to show that $g_1(t) \leq g_s(t)$ for all $0 \leq t \leq T$, and all $s > 0$; then the inequality for $s = 0$ will follow by continuity.

Fixing $s > 0$ and reasoning by contradiction, define T_0 to be the largest value of t_0 such that the conclusion of the theorem holds for all $t \in [0, t_0]$ (recall that s is fixed). That is,

$$T_0 = \sup\{t_0 \colon g_1(t) \leq g_s(t) \text{ for all } t \leq t_0\}.$$

Then obviously $g_1(T_0) = g_s(T_0)$, and $T_0 \neq 0$ since $g_1(0) = g_s(0)$ and $1 = \dot{g}_1(0) > \dot{g}_s(0) = 1 - s$.

Let us show that $\dot{g}_1(T_0) < \dot{g}_s(T_0)$, implying that $g_1(t) \leq g_s(t)$ for all t from some interval $[T_0, T_0 + \varepsilon]$, $\varepsilon > 0$, and thus obtaining a contradiction to the choice of T_0.

To show that $\dot{g}_1(T_0) < \dot{g}_s(T_0)$, one notices that

$$\frac{d^2 g_1}{dt^2}(t) = -K_1(t) g_1(t) \leq -K_0(t) g_0(t) < -(K_0(t) - s) g_s(t) = \frac{d^2 g_s}{dt^2}(t)$$

for all $t \le T_0$ since $g_1(t) \le g_s(t)$ for all $t \le T_0$. Now by integrating the strict inequality

$$\frac{d^2 g_1}{dt^2}(t) < \frac{d^2 g_s}{dt^2}(t),$$

which holds for all t from the *open* interval $]0, T_0[$, and taking into account that $\dot{g}_1(0) = \dot{g}_s(0)$, we get $\dot{g}_1(T_0) < \dot{g}_s(T_0)$. □

Remark. Notice that this proof is purely two-dimensional, for it uses equation (6.15) instead of (6.14); see the remark after equation (6.15).

Remark. The condition that the Jacobi field does not vanish for $t \le T$ is essential. Two points $\gamma(a)$ and $\gamma(b)$ such that there is a nonzero Jacobi field vanishing at both points are said to be *conjugate along* γ. A segment of a geodesic that contains a pair of conjugate points in the interior is *never* a shortest path between its endpoints. (A nice proof uses the index form, which is beyond the scope of this book. A hand-crafted proof is tedious, though elementary.) Hence such segments are of no interest for distance measurements. Conversely, a geodesic segment that does not contain a pair of conjugate points is always a locally shortest path (it is the shortest path among sufficiently close paths). This statement can be easily proven arguing as in the proof of Lemma 5.2.9.

6.5.2. Cartan–Alexandrov–Toponogov Comparison Theorem. Now we are going to exploit a construction which is very similar to the one used in the proof of Lemma 6.4.13. Consider two exponential maps $\exp_p : U_p \to \Omega_p = \exp_p(U_p)$ and $\exp_q : U_q \to \Omega_q = \exp_q(U_q)$, where U_p and U_q are chosen small enough so that both exponential maps are diffeomorphisms.

It is convenient to assume that U_p and U_q are balls of the same radius r. Choose a linear isometry $I : (T_q\Omega_q, \langle,\rangle_q) \to (T_p\Omega_p, \langle,\rangle_p)$. Clearly I sends U_q to U_p. Define $\sigma = \exp_p \circ I \circ \exp_q^{-1}$, $\sigma : \Omega_q \to \Omega_p$.

Alternatively, one can choose normal coordinate systems centered at p and q, and define σ to map every point from Ω_q to the point in Ω_p with the same coordinates.

Let us list some obvious properties of σ. First of all, σ preserves the radial distances:

$$d_{\Omega_q}(q, s) = d_{\Omega_p}(p, \sigma(s)).$$

Indeed, σ maps geodesics starting from q and parameterized by arc length to naturally parameterized geodesics starting from p.

Next, by construction $d_q\sigma$ is a linear isometry between tangent spaces, and hence σ preserves angles between curves starting from q (notice that, since we do not specify whether we mean Riemannian angles or the general notion of an angle between two curves in a length space, we implicitly use Lemma 5.1.14).

Exercise 6.5.2. Reasoning as in Lemmas 6.4.12 and 6.4.13, show that if $K(p) = K(q)$, then σ distorts the metric in no more than the fourth order (as we move away from q).

Exercise 6.5.3. Use Exercise 5.1.19 to show that the length of a circle of radius r centered at p is equal to $\pi r^2(1 - \frac{1}{6}K(p)r^2 + o(r^2))$.

The proof of Lemma 6.4.13 was based on the observation that, for a region Ω of constant curvature k, σ happens to be an isometry. We will generalize this as the following property of σ, which perhaps best explains the geometric meaning of Gaussian curvature:

Theorem 6.5.4. *If the Gaussian curvature of Ω satisfies $k \le K(s)$ for all $s \in \Omega$, then σ is a nonexpanding map in some neighborhood of q; that is, there exists an $\varepsilon > 0$ such that*

$$d(\sigma(s_1), \sigma(s_2)) \le d(s_1, s_2) \quad \textit{for all} \quad s_1, s_2 \in \mathrm{Ball}_\varepsilon(q).$$

If the Gaussian curvature of Ω satisfies $k \ge K(s)$ for all $s \in \Omega$, then σ is a noncontracting map in some neighborhood of q; that is,

$$d(\sigma(s_1), \sigma(s_2)) \ge d(s_1, s_2) \quad \textit{for all} \quad s_1, s_2 \in \mathrm{Ball}_\varepsilon(q).$$

Proof. Let us choose normal coordinates centered at q and p such that σ maps every point in Ω_q to the point in Ω_p with the same coordinates. The Riemannian metrics in Ω and Ω_k are given by given by

$$\langle V, V\rangle = \sqrt{v_x^2 + G(x,y)v_y^2} \quad \text{in} \quad \Omega, \tag{6.18}$$

$$\langle V, V\rangle = \sqrt{v_x^2 + G_k(x,y)v_y^2} \quad \text{in} \quad \Omega_k, \tag{6.19}$$

where G and G_k are the metric coefficients of the metrics in Ω and Ω_k.

As usual, denote by $g(t) = \sqrt{G(t,y)}$ the length of the coordinate vector field Y, which is a Jacobi field along a coordinate line in Ω. Similarly $g_k(t) = \sqrt{G_k(t,y)}$ is the length of the coordinate Jacobi field Y_k along a coordinate line in Ω_k.

Consider the case $k \ge K(s)$ for all $s \in \Omega$; the other one is similar. Applying the Rauch Comparison Theorem 6.5.1 to Y and Y_k, we conclude that $G(x,y) \ge G_k(x,y)$. Now it is obvious from (6.18) and (6.19) that σ does not increase lengths of curves. Now it remains to choose ε small enough so that a shortest curve between any two points in $\mathrm{Ball}_\varepsilon(q)$ is contained in the domain Ω_q of σ (it is enough to choose $\varepsilon = \frac{1}{2}r_q$, where r_q is the injectivity radius at q). □

Exercise 6.5.5. Give another proof of this theorem that uses Lemma 6.4.7 and Exercise 6.4.10 instead of Theorem 6.5.1.

Recall that every point p of a region Ω has a neighborhood U such that there is only one shortest path between any two points in U (Exercise 5.2.11 guarantees this).

Let Ω_k be the Euclidean plane if $k = 0$, the sphere of the Gaussian curvature k if $k > 0$, and the hyperbolic plane of curvature k if $k < 0$.

Let a, b, c be three points in U, and let a', b', c' be three points in Ω_k such that $d(a,c) = d(a',c')$, $d(b,c) = d(b',c')$ and the angle between the shortest paths $[ca]$ and $[cb]$ at c is equal to that between the shortest paths $[c'a']$ and $[c'b']$ at c'.

Together with the fact that σ preserves angles at p, Theorem 6.5.4 immediately implies that a Riemannian metric in Ω enjoys the following version of the Cartan–Alexandrov–Toponogov Comparison Theorem for small triangles:

Theorem 6.5.6. *Let U, Ω_k and triangles $\triangle abc$, $\triangle a'b'c'$ be as above. Then*

$d(a,b) \geq d(a',b')$ provided that the Gaussian curvature in Ω satisfies $k \geq K(s)$ for all $s \in \Omega$, and

$d(a,b) \leq d(a',b')$ if the Gaussian curvature in Ω satisfies $k \leq K(s)$ for all $s \in \Omega$.

Proof. The possibility to choose I in the construction of σ leaves us enough freedom to have σ map a to a' and b to b'. Now the theorem follows from Theorem 6.5.4. □

Exercise 6.5.7. (i) Show that if the Gaussian curvature in a region is nonnegative (nonpositive), then every angle of a triangle is less than or equal to (resp. greater than or equal to) the corresponding angle of a Euclidean triangle with the same lengths of sides.

(ii) Show that if the Gaussian curvature in a region is nonnegative (nonpositive), then the sum of the angles in every triangle is at most (resp. at least) π.

The comparison property given by Theorem 6.5.6 is taken as (one among several) definition of the length spaces of curvature bounded below and above. Hence we have made the necessary preparation for our synthetic machinery by showing that the Riemannian manifolds with bounded sectional curvature are length spaces with the same curvature bounds (in the sense of Chapter 4).

Chapter 7

Space of Metric Spaces

Most of our discussions in this course are about relations between various properties and characteristics of a metric space. In other words, we study one metric space at a time. However it may be useful to consider every particular space as a representative of a class of similar objects, which could be called the "space of metric spaces". There is a similar idea in the foundation of the mathematical analysis: instead of talking about a single number, one studies the entire real line. This brings new notions like continuity, derivatives, etc., and they appear to be powerful tools that give new information about the original objects, numbers. For example, a common method of proving inequalities is finding maxima and minima, and these notions make no sense without considering a set of numbers rather than a single number.

Another example of using this "global" approach is given by the theory of convex sets in $\mathbb{R}^n$. Introducing Hausdorff distance (discussed in Subsection 7.3.1 below) turns the set of all compact convex sets into a metric space. This opens a way to apply "analytic" techniques to convex sets just like numbers. For example, one can use maxima and minima because the space of compact convex sets is boundedly compact just like $\mathbb{R}$ (cf. Theorem 7.3.8). Also, a simple argument shows that polyhedra are dense in this space, and this allows us to extend certain statements "by continuity" from polyhedra to arbitrary convex sets.

In this chapter we extend this approach further and introduce a distance between abstract metric spaces. In fact, we will define several distances suitable for different purposes. Let us make some remarks in advance. First, in most cases the distance itself is not essential; what matters is the topology that it defines. Concerning the "space of metric spaces", we will mainly study converging sequences and similar notions, not exact numerical

values of the distance. Second, to make use of the topological structure one usually needs a collection of quantities that are continuous, properties that are preserved through passing to the limit, etc. The finer the topology is, the more such quantities and properties exist. However, the topology being finer means that there are fewer converging sequences and fewer compact sets, and this makes the topology less useful. For this reason different topologies may be needed. This is illustrated by functional analysis where every natural class of functions comes with its own topology—recall C^0, C^1, L^1, L^2 and so on.

To ignite curiosity, we begin with a mathematical fairy tale, or a science fiction story—for this preparation level. It was a famous open problem whether every group of polynomial growth is virtually nilpotent (since this is a fairy tale, we do not need to define either of the notions). Here is an idea of Gromov's solution of this problem. Gromov suggested considering a sequence of metric spaces (G, hd) where G is our finitely generated group with a word metric d, and h is a positive number tending to zero. Using his criterion for pre-compactness of certain spaces of metric spaces, Gromov specified a subsequence converging to some length space, along with an action of the group on this space (a reader may think of the example $(\mathbb{Z}^2, hd)$, where d is the word metric for the standard choice of two generators, as a sequence of finer and finer grids converging to $\mathbb{R}^2$ with the norm $||(x,y)|| = |x| + |y|$). The limit space happened to be a Finsler manifold. Thus Gromov was able to reduce the conjecture to the case of an isometry group of a manifold, and in this case the answer was to be affirmative.

7.1. Examples

We start with several examples, without formal definitions and proofs. Not all these examples are covered by notions discussed later in this course; some of them will be used for motivations and comparisons only.

7.1.1. Tangent cone of a convex set. Consider a closed convex set in the plane. If this set is essentially two-dimensional, i.e., not contained in a line, its boundary is a continuous curve. Playing with examples one makes the following observations, which are not difficult to formalize and prove. There are two kinds of points on a convex curve. There are points where the curve is "smooth" in the sense that it admits a unique tangent line, and points where the curve "breaks" and has different tangents at two sides. Similarly, a three-dimensional convex body near a point of its boundary may look (up to the first order) like a half-space, or like a dihedral angle with an edge, or like a "sharp" solid angle. Furthermore, a dihedral angle has a certain angular measure, and sharp angles may vary in shape. All these

first-order properties are encoded in the *tangent cone*, which is defined as follows.

Let X be a convex set in $\mathbb{R}^n$ and $p \in X$. For every $\lambda > 0$, consider the homothety with the coefficient λ centered at p. Assuming for convenience that p is the origin of $\mathbb{R}^n$, one obtains the dilated set $\lambda X = \{\lambda x : x \in X\}$. The tangent cone of X at p is by definition the limit of these sets as $\lambda \to \infty$. Since X is convex, it is fairly easy to define what is meant by a limit here. Namely, the family of sets $\{\lambda X\}$ is increasing in the sense that $\lambda_1 X \subset \lambda_2 X$ if $\lambda_1 < \lambda_2$ (this trivially follows from convexity). So these sets naturally "converge" to their union, $\bigcup_{\lambda>0} \lambda X$. It is more convenient to have a closed cone, so one usually takes the closure instead of a bare union. Thus the tangent cone of X at p is defined as the closure of the set $\bigcup_{\lambda>0} \lambda X$, where the homotheties $X \mapsto \lambda X$ are centered at p. It is easy to check that it is indeed a cone (i.e., it consists of entire rays emanating from p, or equivalently is invariant under homotheties). In fact, it is the minimal convex cone (with its vertex at p) that contains X. Note that p can be an internal point of the set; in this case the tangent cone is the entire space. The construction of the tangent cone is pictured in Figure 7.1 (dilated spaces have lighter color).

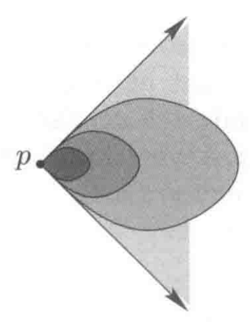


Figure 7.1: Spaces with dilated metrics converge to the tangent cone.

The tangent cone is closely related to the "space of directions" introduced in Subsection 3.6.6. Namely, a ray emanating from p is contained in the tangent cone if and only if its direction belongs to the space of directions at p. This gives an alternative definition for the space of directions. It is more convenient in many cases, because some properties of a set can be automatically transferred from a set to its tangent cone "by continuity". A trivial example is the property of convexity. This kind of definition of a tangent cone (and a space of direction) can be applied to other spaces, not only convex sets; however one needs a suitable definition of limit.

7.1.2. Asymptotic cone. Tangent cones grasp the behavior of a set in a small neighborhood of a point. There is a similarly defined notion of

asymptotic cone that does this "near infinity". Given a (closed) convex set $X \subset \mathbb{R}^n$ and a point $p \in X$, the asymptotic cone is the limit of homothetic sets λX as $\lambda \to 0$ (the homotheties are centered at p). As in the case of tangent cone, there is no problem with a definition of the limit. Since $\{\lambda X\}$ is a nested family of sets, the "limit" is just their intersection, $\bigcap_{\lambda>0} \lambda X$ (see Figure 7.2). It is easy to see that the asymptotic cone is just the union of all rays emanating from p and contained in X. The shape of the asymptotic cone does not depend on the choice of p in the sense that asymptotic cones with different choices for p are parallel translations of one another.

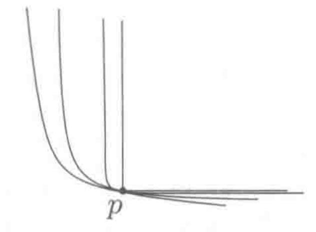


Figure 7.2: Spaces with dilated metrics converge to the asymptotic cone.

Loosely speaking, the asymptotic cone is what one sees observing the set from far away (assuming that moving the observation point stretches the visual image homothetically). A similar association applies to tangent cones: the tangent cone at p is what an observer sees looking at p under a microscope. The surface of the Earth looks planar to us; indeed, the tangent cone of the sphere is a plane (more precisely, that of a solid ball is a half-space).

We will extend the above definition of the asymptotic cone to general metric spaces in the next chapter (Section 8.2).

7.1.3. Asymptotic cone of a lattice. Consider the group $\mathbb{Z}^2$ which can be thought of as the lattice of integer points in $\mathbb{R}^2$. In Subsection 3.2.3 we described how to define a word metric for any given set of generators. Namely, the word metric is the intrinsic metric of the group's Cayley graph. (More precisely, it is the restriction of that metric to the group, but we prefer to keep the graph in this example.) Take the simplest possible set of generators, namely two generators $(1,0)$ and $(0,1)$. Then the Cayley graph is naturally identified with the grid of horizontal and vertical lines passing through entire points in the plane. The intrinsic metric of the graph coincides with the one induced from the plane.

Let's try to apply (informally) the definition of the asymptotic cone to this graph. Let the graph be denoted by X. The dilated space λX is the grid of all vertical and horizontal lines whose coordinates are integer multiples of

λ. In other words, these lines divide $\mathbb{R}^2$ into squares of side λ. As λ goes to zero, the grid becomes finer and finer, and it is natural to think that its limit should include all points in the plane. (This is only a speculation appealing to the reader's visual intuition. We don't have a suitable definition of limit yet.)

However, the limit as a metric space is not the Euclidean plane $\mathbb{R}^2$ because the distances in the grids are essentially different from the Euclidean one. In fact, the limit is the normed space $\mathbb{R}^2_1$ which is $\mathbb{R}^2$ with the norm $\|(x,y)\|_1 = |x| + |y|$. Indeed, looking at the shortest paths in the grid λX one can see that the distance between two nodes (x_1, y_1) and (x_2, y_2) equals $|x_1 - x_2| + |y_1 - y_2|$. It is the same as the distance in $\mathbb{R}^2_1$. And for a small λ the distance between nonnodes is only slightly different (by no more than 4λ) from the distance in $\mathbb{R}^2_1$ (because for every point in the grid there is a node within a distance $\le \lambda$ from it). Thus if we want the distances in the grid to converge (in any sense) to the distances in the limit space, we have to define the latter as $\mathbb{R}^2_1$.

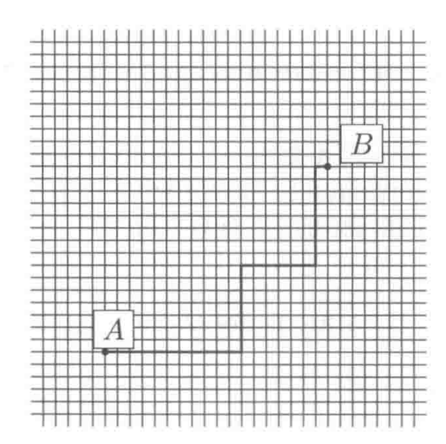


Figure 7.3: A fine grid and a shortest path.

In the above considerations we used an isometric embedding of the Cayley graph into $\mathbb{R}^2$. Such an embedding may fail to exist for other sets of generators, but a reasonable asymptotic cone can be defined for every word metric. In fact, the asymptotic cone is always a two-dimensional normed space. Its unit ball can be identified with a polygon in $\mathbb{R}^2$ obtained as a convex hull of the set of generators and vectors opposite to generators. We will discuss this further in Subsection 8.5.1.

7.1.4. Converging surfaces. Smooth surfaces in $\mathbb{R}^3$ can be (at least locally) parameterized by a domain $D \subset \mathbb{R}^2$, either as graphs of smooth functions from D to R, or as images of embeddings from D to $\mathbb{R}^3$. Hence a

topology on a set of surfaces can be derived from a topology on the set of maps. Namely, a sequence of surfaces converges if a sequence of their suitable parameterizations do. Of course, if a surface is covered by several patches, each with its own parameterizations, certain compatibility conditions should be imposed. Anyway, one can speak, for instance, about C^k-convergence of surfaces ($k = 0, 1, \dots$).

Remark 7.1.1. If some characteristic of a surface is defined in terms of derivatives up to the kth order, one usually has to choose at least the C^k topology in order to make the characteristic continuous. However, imposing geometric restrictions sometimes allows one to use a weaker topology.

For example, consider regular smooth curves in the plane. The length and curvature (which are defined in terms of first and second derivatives, respectively) are not continuous with respect to C^0 topology. The length is lower semi-continuous, and the curvature is not semi-continuous from either side. But if the class of curves is restricted to *convex* ones (a planar curve is called convex if it is contained in a boundary of a convex set), the situation improves. Namely on the set of convex regular curves equipped with C^0 topology, the length is continuous and the minimum of curvature is upper semi-continuous.

This is a partial case of a general phenomenon: convexity and lower curvature bounds are surprisingly stable with respect to, say, passing to a limit in a relatively weak topology (like the C^0 one in the example with convex curves).

Exercise 7.1.2. Prove the statements about regular curves from the above remark.

A "functional" convergence discussed above is hardly suitable for our purposes since we mainly deal with abstract metric spaces rather than subsets of $\mathbb{R}^n$. One can modify this notion using the following idea. If two surfaces are parameterized by the same domain, the two parameterizations determine a one-to-one correspondence between points of the surfaces. In other words, they define a homeomorphism from one of them to the other. If the surfaces are close enough to each other, this homeomorphism only slightly changes certain data such as distances, metric tensors, or their derivatives. One can turn this to a definition saying that two spaces have small distance between them if there is a homeomorphism from one to the other which "almost preserves" certain geometric characteristics, for example the distance. Later we will give several precise definitions of this sort.

This approach has a disadvantage: it requires spaces to be homeomorphic. Sometimes this requirement is too restrictive. Let X be the standard

two-dimensional sphere and X_n be same sphere with a small handle attached to it. (Let the diameter of the handle be less than $1/n$.) As n grows, handles become smaller and smaller and spaces X_n look more and more similar to X (see Figure 7.4). One could say that handles vanish to a point and thus X_n converge to X. However X_n is not homeomorphic to X, so a different notion of convergence is needed here.

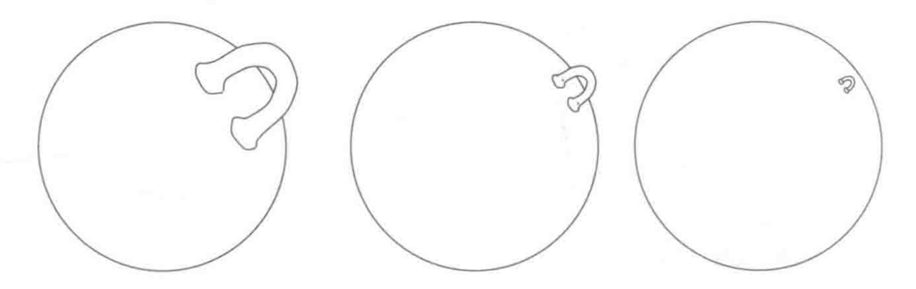


Figure 7.4: Spheres with vanishing handles.

Furthermore, even for two homeomorphic and similarly looking spaces, it may happen that the "similarity" is not realized by a homeomorphism. For example, consider two spheres of essentially different radii connected by a long thin tube. One can attach a tiny little handle to either the larger or the smaller sphere; let X and Y be the resulting spaces (see Figure 7.5). Then X and Y are homeomorphic, but any homeomorphism (in fact, any continuous map) from X to Y essentially distorts distances between some points, no matter how small the handles are. Nevertheless X and Y are "close" to each other in the same sense as in the previous example.

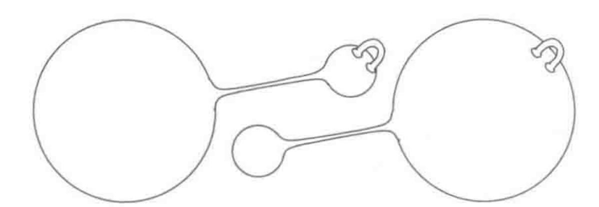


Figure 7.5: A homeomorphism has to change distances essentially.

The reader already familiar with Hausdorff distance may notice that it provides some way to formalize the above examples: the distance between these sets is small because each of them is contained in a small neighborhood of the other in $\mathbb{R}^3$. However, if we consider two surfaces as length spaces with close intrinsic metrics, and it is not at all clear that these metrics can be realized by close embedding (even if they can be realized by embeddings

at all). Intrinsic geometry of a surfaces is in a complicated relation with its "visual image" in $\mathbb{R}^3$ and measuring the distance in $\mathbb{R}^3$ makes little sense. Nevertheless, measuring the distance in an ambient space is a step in the right direction. Later in this chapter we will introduce a very important Gromov–Hausdorff distance between metric spaces. As the name suggests, this notion is derived from the classic Hausdorff distance.

7.1.5. Uniform convergence. Recall that a sequence $\{f_n\}$ of real-valued functions on a set X is said to *uniformly converge* to a function f if

$$\sup_{x\in X} |f_n(x) - f(x)| \to 0$$

as $n \to \infty$. Since every metric on X is a real-valued function (defined on $X \times X$), the notion of uniform convergence applies to metrics: a sequence $\{d_n\}$ of metrics on X uniformly converges to a metric d (or, $d_n \rightrightarrows d$) if

$$\sup_{x,x'\in X} |d_n(x,x') - d(x,x')| \to 0$$

as $n \to 0$. We have already seen (Exercise 2.4.19) that passing to a uniform limit of metrics preserves the property of being intrinsic: if the metrics d_n are intrinsic, the limit metric d is intrinsic too. Nevertheless, uniform convergence is a relatively weak type of convergence. For example, a uniform limit of Riemannian metric can be non-Riemannian.

Exercise 7.1.3. Let D^2 denote the standard unit ball in $\mathbb{R}^2$. Prove that the metric of $\mathbb{R}^2_1$ (that is, $\mathbb{R}^2$ with the norm $\|(x,y)\| = |x| + |y|$) restricted to D^2 can be obtained as a uniform limit of Riemannian metrics.

Hint: Fill the cells in the fine grid from Subsection 7.1.3.

In fact, *any* intrinsic metric on D^2 can be obtained as a uniform limit of Riemannian metrics. In view of this, it seems unbelievable that a uniform convergence preserves some information about the curvature. However it does: if the curvatures of converging metrics are uniformly bounded from below, this curvature bound (in the Alexandrov sense) is inherited by the limit metric! In fact, this holds for a more general kind of convergence, namely, the Gromov–Hausdorff one. We will prove this later in the course.

The uniform convergence of metrics has one unnatural restriction: the metrics should be defined on the same set. Taking into account the underlying set on which a metric is defined would make us distinguish metric spaces that are otherwise identical (i.e., isometric). To avoid this, one could introduce the following (preliminary) definition of uniform convergence of metric spaces:
indexconvergence!uniform a sequence $\{X_n\}$ of metric spaces uniformly converges to a metric space (X,d) if there is a sequence $\{d_n\}$ of metrics on X such that (X,d_n) is isometric to X_n for all n, and $d_n \rightrightarrows d$.

This definition is "preliminary" because it can be reformulated in a more convenient way. Our formulation, when saying that (X, d_n) is isometric to X_n, implicitly includes an isometry map from X_n to (X, d_n). Constructing an isometry for two metrics (that are known to be isometric) may be difficult, whereas the metric d_n can be trivially recovered from such an isometry map. This observation suggests that our definition would be more usable if it dealt with maps from X_n to X rather than with metrics on X.

Definition 7.1.4. Let X and Y be metric spaces and $f : X \to Y$ an arbitrary map. The *distortion* of f is defined by

$$\operatorname{dis} f = \sup_{x_1, x_2 \in X} |d_Y(f(x_1), f(x_2)) - d_X(x_1, x_2)|$$

where d_X and d_Y are the metrics of X and Y.

The definition of distortion resembles that of dilatation of a Lipschitz map. The only difference is that the latter measures *relative* change of distances while the former measures *absolute* ones. The notion of distortion can be applied to noncontinuous maps.

Definition 7.1.5. We say that a sequence $\{X_n\}$ of metric spaces *uniformly converges* to a metric space X if there exist homeomorphisms $f_n : X_n \to X$ such that $\operatorname{dis}(f_n) \to 0$ as $n \to \infty$.

Exercise 7.1.6. Prove that Definition 7.1.5 is equivalent to our "preliminary" definition of uniform convergence.

Exercise 7.1.7. Define a distance between metric spaces such that the convergence with respect to this distance is equivalent to the uniform convergence in the sense of Definition 7.1.5.

Hint: See the definition of Lipschitz distance in the next section.

7.2. Lipschitz Distance

The idea of Lipschitz distance is the following: two metric spaces X and Y are close to each other if there is a homeomorphism $f: X \to Y$ such that the ratios $d_Y(f(x), f(x'))/d_X(x, x')$ are close to 1. In other words, Lipschitz distance measures *relative* difference between metrics. Note that relative differences and relative errors is just what people care about concerning the metric of the physical universe. It is a good achievement to find the distance between the Sun and the Earth with an error of about a thousand miles, but measuring one's apartment with such a precision is not a good idea.

Let X and Y be two metric spaces. Recall that the dilatation of a Lipschitz map $f: X \to Y$ is defined by

$$\operatorname{dil} f = \sup_{x, x' \in X} \frac{d_Y(f(x), f(x'))}{d_X(x, x')}$$

where d_X and d_Y are the metrics of X and Y. A homeomorphism f is called bi-Lipschitz if both f and f^{-1} are Lipschitz maps.

Definition 7.2.1. The *Lipschitz distance* d_L between two metric spaces X and Y is defined by

$$d_L(X,Y) = \inf_{f\colon X\to Y} \log\bigl(\max\{\operatorname{dil}(f), \operatorname{dil}(f^{-1})\}\bigr)$$

where the infimum is taken over all bi-Lipschitz homeomorphisms $f\colon X \to Y$.

A sequence $\{X_n\}_{n=1}^{\infty}$ of metric spaces is said to *converge* in the Lipschitz sense to a metric space X if $d_L(X_n, X) \to 0$ as $n \to \infty$.

If there are no bi-Lipschitz homeomorphisms from X to Y, then one sets $d_L(X,Y) = \infty$. Thus the Lipschitz distance is not suitable for comparing metric spaces that are not (bi-Lipschitz) homeomorphic.

Example 7.2.2. Let M be a smooth manifold and $\{F_n\}_{n=1}^{\infty}$ be a sequence of Finslerian structures on M. Assume that it converges to a Finslerian structure F in the sense that $F_n(v)/F(v) \to 1$ as $n \to \infty$ uniformly over all nonzero vectors $v \in TM$ (recall that Finslerian structures are functions on TM). Then $d_L((M,d_n),(M,d)) \to 0$ where d_n and d are the Finslerian metrics corresponding to F_n and F.

Example 7.2.3. Let $\{M_t\}_{t\in\mathbb{R}}$ be a family of smooth surfaces in $\mathbb{R}^3$ parameterized by maps $f_t\colon \Omega \to \mathbb{R}^3$ where Ω is a region in $\mathbb{R}^2$. Assume that the family itself is smooth (at least C^1); i.e., the function $F\colon \Omega \times \mathbb{R} \to \mathbb{R}^3$ defined by $F(x,t) = f_t(x)$ is smooth. Then $d_L(M_t, M_0) \to 0$ as $t \to 0$. This conclusion may fail if the family is only continuous.

Theorem 7.2.4. *d_L is nonnegative, symmetric and satisfies the triangle inequality. For compact spaces X and Y, $d_L(X,Y) = 0$ if and only if X and Y are isometric.*

Proof. If $f\colon X \to Y$ is a homeomorphism, then at least one of the dilatations $\operatorname{dil} f$ and $\operatorname{dil} f^{-1}$ is greater than or equal to 1 because f and f^{-1} cannot both decrease the distances. Hence $d_L(X,Y) \ge 0$. The symmetry of d_L is trivial.

To prove the triangle inequality, let X, Y, Z be metric spaces and $f\colon X \to Y$ and $g\colon Y \to Z$ be bi-Lipschitz homeomorphisms. Then $h = g \circ f$ is a bi-Lipschitz homeomorphism from X to Z. Moreover $\operatorname{dil}(h) \le \operatorname{dil}(f) \cdot \operatorname{dil}(g)$; hence $\log(\operatorname{dil} h) \le \log(\operatorname{dil} f) + \log(\operatorname{dil} g)$. Together with the similar inequality for $h^{-1} = f^{-1} \circ g^{-1}$, this implies the triangle inequality $d_L(X,Z) \le d_L(X,Y) + d_L(Y,Z)$.

If X and Y are isometric, substituting an isometry $f\colon X \to Y$ in the definition of d_L shows that $d_L(X,Y) = 0$. The converse statement is more

delicate. By the definition of d_L, the relation $d_L(X,Y) = 0$ implies that there exists a sequence of maps $f_n \colon X \to Y$, $n \in \mathbb{N}$, such that $\operatorname{dil} f_n \to 1$ and $\operatorname{dil}(f^{-1}) \to 1$ as $n \to \infty$. Since the dilatations of f_n are uniformly bounded, by the Arzela–Ascoli Theorem a subsequence of $\{f_n\}$ uniformly converges to a map $f \colon X \to Y$. We may assume (without loss of generality) that the sequence f_n itself converges to f. Since $\operatorname{dil} f_n \to 1$ as $n \to \infty$, for all $x, x' \in X$ one has $|f_n(x)f_n(x')|/|xx'| \to 1$ and hence $|f(x)f(x')| = |xx'|$. Thus f is a distance-preserving map. Similarly, there is a distance-preserving map $g : Y \to X$. The composition $f \circ g$ is a distance-preserving map from Y to itself. Since Y is compact, the map $f \circ g$ is bijective by Theorem 1.6.14. Hence f is surjective and therefore is an isometry. □

This theorem tells us that the Lipschitz distance is a metric on the "space" of isometry classes of compact metric spaces. This "space" is not a good object from the rigorous point of view of the Set Theory (just like the "set of all sets"). However we will refer to this space in our formulations. For us, this space is just a collection of elements (which are isometry classes of metric spaces) and not an object of the Set Theory, so no set-theoretic paradoxes can occur. Furthermore, all our statements about this "space of metric spaces" can be reformulated in terms of its elements (like the above theorem), so one can think of this space as a way to shorten formulations.

Remark 7.2.5. Alternatively, one can formally justify "the space of compact metric spaces" using the following observation: this "space" is of cardinality continuum; i.e., there can be no more than continuum of mutually nonisometric compact spaces. (Exercise: prove this.) Therefore, choosing a representative from each isometry class one obtains a collection of cardinality continuum, which is a legitimate set.

Exercise 7.2.6. Prove that Lipschitz convergence of compact metric spaces implies uniform convergence (Definition 7.1.5).

Exercise 7.2.7. Prove that Lipschitz convergence is equivalent to the uniform convergence within the class of *finite* metric spaces (that is, consisting of finitely many points).

Remark 7.2.8. The two types of convergence are not equivalent in general. There are sequences of compact metric spaces that converge uniformly but do not converge w.r.t. the Lipschitz distance (can you give an example?)

7.3. Gromov–Hausdorff Distance

The Gromov–Hausdorff distance is similar to the Lipschitz distance in the sense that it is a distance between compact metric spaces, where the latter ones are considered up to an isometry. However it determines a

weaker "topology"; in particular, the distance is always finite (and can be arbitrarily small) even for nonhomeomorphic spaces. The difference between the Lipschitz and the Gromov–Hausdorff distances is somewhat similar to the one between C^1 and C^0 norms in functional spaces.

7.3.1. Hausdorff distance. First we introduce the ordinary Hausdorff distance. This is a distance between subsets of a metric space, not between abstract metric spaces.

We denote by $U_r(S)$ the r-neighborhood of a set S in a metric space, i.e., the set of points x such that $\text{dist}(x, S) < r$. Equivalently, $U_r(S) = \bigcup_{x \in S} B_r(s)$.

Definition 7.3.1. Let A and B be subsets of a metric space. The *Hausdorff distance* between A and B, denoted by $d_H(A, B)$, is defined by

$$d_H(A, B) = \inf\{r > 0 : A \subset U_r(B) \text{ and } B \subset U_r(A)\}.$$

The next exercise contains convenient reformulations of the definition.

Exercise 7.3.2. Let A and B be subsets of a metric space and $r > 0$.

1. Prove that $d_H(A, B) = \max\{\sup_{a \in A} \text{dist}(a, B),\ \sup_{b \in B} \text{dist}(b, A)\}$.

2. Prove that $d_H(A, B) \le r$ if and only if $\text{dist}(a, B) \le r$ for all $a \in A$ and $\text{dist}(b, A) \le r$ for all $b \in B$.

3. Show that the previous statement would fail if one replaces $\le$ by $<$.

Proposition 7.3.3. *Let X be a metric space. Then*

(1) *d_H is a semi-metric on 2^X (the set of all subsets of X).*

(2) *$d_H(A, \overline{A}) = 0$ for any $A \subset X$ where $\overline{A}$ denotes the closure of A.*

(3) *If A and B are closed subsets of X and $d_H(A, B) = 0$, then $A = B$.*

Proof. 1. Nonnegativity and symmetry are obvious. The triangle inequality follows from the following fact: for any $A \subset X$ and any $r_1, r_2 > 0$ one has $U_{r_1+r_2}(A) \subset U_{r_1}(U_{r_2}(A))$. This is in turn an immediate consequence of the triangle inequality in X.

2. $\text{dist}(x, \overline{A}) = 0$ for all $x \in A$ because $S \subset \overline{A}$. For an $x \in \overline{A}$ one has $\text{dist}(x, A) = 0$ by the definition of closure. Hence $d_H(A, \overline{A}) = 0$.

3. Suppose the contrary, e.g., there is an $x \in A \setminus B$. Since B is a closed set, there is an $r > 0$ such that the ball $B_r(x)$ does not intersect B. Then $x \notin U_r(B)$; hence $d_H(A, B) \ge r > 0$. □

Let $\mathfrak{M}(X)$ denote the set of closed subsets of X equipped with Hausdorff distance. The above proposition tells us that the $\mathfrak{M}(X)$ is a metric space. Moreover every element of the quotient $2^X/d_H$ can be represented by a closed set and therefore $2^X/d_H$ is naturally identified with $\mathfrak{M}(X)$.

Exercise 7.3.4. Let a sequence of sets $A_i \in \mathfrak{M}(X)$ converge to a set $A \in \mathfrak{M}(X)$ with respect to the Hausdorff distance (in short: $A_i \to A$ in $\mathfrak{M}(X)$). Prove that

1. A is the set of limits of all converging sequences $\{a_n\}$ in X such that $a_n \in A_n$ for all n.

2. $A = \bigcap_{n=1}^{\infty}(\text{closure of } \bigcup_{m=n}^{\infty} A_m)$.

Exercise 7.3.5. Let X be a compact metric space and $\{A_i\}_{i=1}^{\infty}$ a sequence of its compact subspaces. Prove that:

1. If $A_{i+1} \subset A_i$ for all i, then $\{A_i\}$ converges in $\mathfrak{M}(X)$ to the intersection $\bigcap_i A_i$.

2. If $A_i \subset A_{i+1}$ for all i, then $\{A_i\}$ converges in $\mathfrak{M}(X)$ to the closure of the union $\bigcup_i X_i$.

Exercise 7.3.6. Let $A_i \to A$ in $\mathfrak{M}(\mathbb{R}^n)$ and all sets A_i are convex. Prove that A is convex. In other words, the set of compact convex sets is closed in $\mathfrak{M}(\mathbb{R}^n)$.

Proposition 7.3.7. *If X is complete, then $\mathfrak{M}(X)$ is complete.*

Proof. Let $\{S_n\}_{n=1}^{\infty}$ be a Cauchy sequence in $\mathfrak{M}(X)$. Let S denote the set of all points $x \in X$ such that for any neighborhood U of x one has $U \cap S_n \neq \emptyset$ for infinitely many n. We will prove that $S_n \to S$. Fix $\varepsilon > 0$ and let n_0 be such that $d_H(S_n, S_m) < \varepsilon$ for all $m, n \geq n_0$. It suffices to show that $d_H(S, S_n) < 2\varepsilon$ for any $n \geq n_0$.

1. $\operatorname{dist}(x, S_n) < 2\varepsilon$ for an $x \in S$. There exists an $m \geq n_0$ such that $B_\varepsilon(x) \cap S_m \neq \emptyset$. In other words, there is a point $y \in S_m$ such that $|xy| < \varepsilon$. Since $d_H(S_m, S_n) < \varepsilon$, one has $\operatorname{dist}(y, S_n) < \varepsilon$ and therefore $\operatorname{dist}(x, S_n) \leq |xy| + \operatorname{dist}(y, S_n) < 2\varepsilon$.

2. $\operatorname{dist}(x, S) < 2\varepsilon$ for an $x \in S_n$. Let $n_1 = n$ and for every integer $k > 1$ chose an index n_k such that $n_k > n_{k+1}$ and $d_H(S_p, S_q) < \varepsilon/2^k$ for all $p, q \geq n_k$. Then define a sequence of points $\{x_k\}$, where $x_k \in S_{n_k}$, as follows. Let $x_1 = x$, and x_{k+1} be a point of $S_{n_{k+1}}$ such that $|x_k x_{k+1}| < \varepsilon/2^k$ for all k. Such a point can be found because $d_H(S_{n_k}, S_{n_{k+1}}) < \varepsilon/2^k$. Since $\sum_{k=1}^{\infty} |x_k x_{k+1}| < 2\varepsilon < \infty$, the sequence $\{x_k\}$ is a Cauchy sequence and hence it converges to a point $y \in X$. Then $|xy| = \lim |x x_n| \leq \sum |x_k x_{k+1}| < 2\varepsilon$. Since $y \in S$ by construction, it follows that $\operatorname{dist}(x, S) < 2\varepsilon$. □

Theorem 7.3.8 (Blaschke). *If X is compact, then $\mathfrak{M}(X)$ is compact.*

Proof. By the previous proposition, $\mathfrak{M}(X)$ is complete. Therefore it suffices to prove that $\mathfrak{M}(X)$ is totally bounded. Let S be a finite ε-net in X. We will prove that 2^S is an ε-net in $\mathfrak{M}(X)$. Let $A \in \mathfrak{M}(X)$. Consider the set

$S_A \in 2^S$ defined by

$$S_A = \{x \in S : \text{dist}(x, A) \le \varepsilon\}.$$

Since S is an ε-net in X, for every $y \in A$ there exists an $x \in S$ such that $|xy| \le \varepsilon$. Since $\text{dist}(x, A) \le |xy| \le \varepsilon$, this point x belongs to S_A. Therefore $\text{dist}(y, S_A) \le \varepsilon$ for all $y \in A$. Since $\text{dist}(x, A) \le \varepsilon$ for any $x \in S_A$ (by the definition of S_A), it follows that $d_H(A, S_A) \le \varepsilon$. Since A is arbitrary, this proves that 2^S is an ε-net in $\mathfrak{M}(X)$. □

Remark 7.3.9. In the theory of convex sets, the following statement is known as Blaschke's Theorem: the set of all compact *convex* subsets contained in any fixed closed ball in $\mathbb{R}^n$ is compact with respect to the Hausdorff distance. This follows from Theorem 7.3.8 and Exercise 7.3.6.

7.3.2. Gromov–Hausdorff distance. Now we are in position to define the Gromov–Hausdorff distance between metric spaces. The idea behind the definition is the following. First, the distance between subspaces in the same metric space is no greater than the Hausdorff distance between them. In other words, if two subspaces of the same space are close to each other in the sense of Hausdorff distance in the ambient space, they must be close to each other as abstract metric spaces. Second, one definitely wants the distance between isometric spaces to be zero. The Gromov–Hausdorff distance is in fact the maximum distance satisfying these two requirements.

Definition 7.3.10. Let X and Y be metric spaces. The *Gromov–Hausdorff distance* between them, denoted by $d_{GH}(X, Y)$, is defined by the following relation. For an $r > 0$, $d_{GH}(X, Y) < r$ if and only if there exist a metric space Z and subspaces X' and Y' of it which are isometric to X and Y respectively and such that $d_H(X', Y') < r$. In other words, $d_{GH}(X, Y)$ is the infimum of positive r for which the above Z, X' and Y' exist. Here d_H denotes the Hausdorff distance between subsets of Z.

Note that X' and Y' in the above definition are regarded with the restriction of the metric of the ambient space Z (as opposed to the induced intrinsic metric). For example, if X is a sphere with its standard Riemannian metric, one cannot take $Z = \mathbb{R}^3$ and $X' = S^2 \subset \mathbb{R}^3$ because X and X' would be only path-isometric but not isometric.

It is trivial that the Gromov–Hausdorff distance between isometric spaces is zero. Later we will show that d_{GH} is a metric in the same sense as d_L; i.e., it is a metric on the space of the isometry classes of compact metric spaces (Theorem 7.3.30). Naturally one says that a sequence $\{X_n\}_{n=1}^{\infty}$ of (compact) metric spaces converges in the Gromov–Hausdorff sense to a (compact) metric space X if $d_{GH}(X_n, X) \to 0$. For noncompact spaces

a slightly more general notion of convergence is used; we will define it in Section 8.1.

Example 7.3.11. If Y is an ε-net in a metric space X, then $d_{GH}(X,Y) \le \varepsilon$. Indeed, one can take $Z = X' = X$ and $Y' = Y$.

Remark 7.3.12. Definition 7.3.10 deals with a huge class of metric spaces, namely, all metric spaces Z that contain subspaces isometric to X and Y. It is possible to reduce this class to disjoint unions of X and Y. More precisely, the Gromov–Hausdorff distance between two metric spaces (X, d_X) and (Y, d_Y) is the infimum of $r > 0$ such that there exists a (semi-)metric d on the disjoint union $X \sqcup Y$ such that the restrictions of d to X and Y coincide with d_X and d_Y and $d_H(X,Y) < r$ in the space $(X \sqcup Y, d)$. In other words, $d_{GH}(X,Y) = \inf\{d_H(X,Y)\}$, where the infimum is taken over all (semi-)metrics on $X \sqcup Y$ extending the ones of X and Y.

To prove this, simply identify $X \sqcup Y$ with $X' \cup Y' \subset Z$ (the notation is from Definition 7.3.10). More formally, fix isometries $f\colon X \to X'$ and $g : Y \to Y'$, then define the distance between $x \in X$ and $y \in Y$ by $d(x,y) = d_Z(f(x), g(y))$. This yields a semi-metric on $X \sqcup Y$ for which $d_H(X,Y) < r$ (if $X' \cap Y' \neq \emptyset$, it may happen that $d(x,y) = 0$). The quotient metric space $(X \sqcup Y)/d$ is isometric to $X' \cup Y'$. To obtain a metric (not a semi-metric) on $X \sqcup Y$, define $d(x,y) = d_Z(f(x), g(y)) + \delta$ where δ is an arbitrary positive constant. Then $d_H(X,Y) < r + \delta$.

Exercise 7.3.13. Prove that $d_{GH}(X,Y) < \infty$ if X and Y are bounded metric spaces.

Hint: Choose a sufficiently large constant $C > 0$ and set $d(x,y) = C$ for all $x \in X$ and $y \in Y$.

Exercise 7.3.14. Let X and Y be two metric spaces and $\operatorname{diam} X < \infty$. Prove that $d_{GH}(X,Y) \ge \frac{1}{2}|\operatorname{diam} X - \operatorname{diam} Y|$.

In other words, the diameter, as a function of a metric space, is a Lipschitz function with a Lipschitz constant 2.

Exercise 7.3.15. Let P be a metric space consisting of one point. Prove that $d_{GH}(X,P) = \operatorname{diam}(X)/2$ for any metric space X.

Proposition 7.3.16. *d_{GH} satisfies the triangle inequality, i.e.,*

$$d_{GH}(X_1, X_3) \le d_{GH}(X_1, X_2) + d_{GH}(X_2, X_3)$$

for any metric spaces X_1, X_2, X_3.

Proof. Let d_{12} and d_{23} be metrics on $X_1 \sqcup X_2$ and $X_2 \sqcup X_3$, respectively, that extend metrics of X_1, X_2 and X_3. Define the distance between $x_1 \in X_1$ and $x_3 \in X_3$ by $d_{13}(x_1, x_3) = \inf_{x_2 \in X_2}\{d_{12}(x_1, x_2) + d_{23}(x_2, x_3)\}$. It is easy

to check (exercise!) that d_{13}, together with the metrics of X_1 and X_3, satisfies the triangle inequality and hence is a metric on $X_1 \cup X_3$. The definition of d_{13} yields that $d_H(X_1, X_3) \le d_H(X_1, X_2) + d_H(X_2, X_3)$, where $d_H(X_i, X_j)$ is taken with respect to the metric d_{ij} $(i, j = 1, 2, 3)$. Taking the infimum over all metrics d_{12} and d_{23} we obtain the desired inequality $d_{GH}(X_1, X_3) \le d_{GH}(X_1, X_2) + d_{GH}(X_2, X_3)$. □

7.3.3. Reformulations. A direct application of the definition of Gromov–Hausdorff distance requires constructing a new metric space Z (or a metric on $X \cup Y$) and verifying the triangle inequality. As the reader could observe, this involves cumbersome details even in simple cases. It would be more convenient if we could compute or estimate $d_{GH}(X, Y)$ by comparing the distances within X and Y to each other, as we did in the cases of uniform and Lipschitz distance. We will give several criteria of this sort. The first and the most general one is based on the notion of *correspondence*.

Roughly speaking, having a correspondence between two metric spaces (or just sets) X and Y means that for every point of X there are one or more "corresponding" points in Y, and vice versa. The criterion that we are going to prove is the following: $d_{GH}(X, Y) < r$ if and only if there is a correspondence between X and Y such that if $x, x' \in X$ and $y, y' \in Y$ are corresponding pairs of points, then $|d_X(x, x') - d_Y(y, y')| < 2r$. The precise definitions and formulations follow.

Definition 7.3.17. Let X and Y be two sets. A *correspondence* between X and Y is a set $\mathfrak{R} \subset X \times Y$ satisfying the following condition: for every $x \in X$ there exists at least one $y \in Y$ such that $(x, y) \in \mathfrak{R}$, and similarly for every $y \in Y$ there exists an $x \in X$ such that $(x, y) \in \mathfrak{R}$.

In other words, a correspondence is just a relation between points of X and Y such that every point of X and Y is in the relation to at least one point in the other set. If $\mathfrak{R}$ is clear from the context, we will say that x and y "correspond" to each other instead of writing $(x, y) \in \mathfrak{R}$.

Example 7.3.18. Any surjective map $f : X \to Y$ be defines a correspondence $\mathfrak{R}$ between X and Y given by

$$\mathfrak{R} = \{(x, f(x)) : x \in X\}.$$

We will call this $\mathfrak{R}$ the *correspondence associated with* f.

Not every correspondence is associated with a map. A closer analog of a correspondence is a "multiple-valued" map, for which a single point is allowed to have more than one image. We will not discuss multiple-valued maps. There is an alternative method, given in the next example, to define a correspondence by means of (ordinary) maps.

Example 7.3.19. Let $f : Z \to X$ and $g : Z \to Y$ be two surjective maps from some set Z. Then a correspondence $\mathfrak{R}$ can be defined by

$$\mathfrak{R} = \{(f(z), g(z)) : z \in Z\}.$$

Exercise 7.3.20. Prove that any correspondence can be obtained via the construction from Example 7.3.19.

Definition 7.3.21. Let $\mathfrak{R}$ be a correspondence between metric spaces X and Y. The *distortion* of $\mathfrak{R}$ is defined by

$$\operatorname{dis}\mathfrak{R} = \sup\{|d_X(x, x') - d_Y(y, y')| : (x, y), (x', y') \in \mathfrak{R}\}$$

where d_X and d_Y are the metrics of X and Y respectively.

Exercise 7.3.22. Prove that for a correspondence $\mathfrak{R}$ associated with a map $f : X \to Y$ as in Example 7.3.18, one has $\operatorname{dis}\mathfrak{R} = \operatorname{dis} f$. (The distortion of a map was introduced in Definition 7.1.4.)

Exercise 7.3.23. Let $\mathfrak{R}$ be obtained from maps $f : Z \to X$ and $g : Z \to Y$ as in Example 7.3.19. Prove that

$$\operatorname{dis}\mathfrak{R} = \sup_{z,z' \in Z} |d_X(f(z), f(z')) - d_Y(g(z), g(z'))|.$$

Exercise 7.3.24. Let $\mathfrak{R}$ be a correspondence between metric spaces X and Y. Prove that $\operatorname{dis}\mathfrak{R} = 0$ if and only if $\mathfrak{R}$ is associated with an isometry map from X to Y.

Theorem 7.3.25. *For any two metric spaces X and Y,*

$$d_{GH}(X, Y) = \frac{1}{2} \inf_{\mathfrak{R}}(\operatorname{dis}\mathfrak{R})$$

where the infimum is taken over all correspondences $\mathfrak{R}$ between X and Y.

In other words, $d_{GH}(X, Y)$ is equal to the infimum of $r > 0$ for which there exists a correspondence between X and Y with $\operatorname{dis}\mathfrak{R} < 2r$.

Proof. 1. For any $r > d_{GH}(X, Y)$, there exists a correspondence $\mathfrak{R}$ with $\operatorname{dis}\mathfrak{R} < 2r$. Indeed, since $d_{GH}(X, Y) < r$, we may assume that X and Y are subspaces of some metric space Z and $d_H(X, Y) < r$ in Z. Define

$$\mathfrak{R} = \{(x, y) : x \in X,\ y \in Y,\ d(x, y) < r\}$$

where d is the metric of Z. That $\mathfrak{R}$ is a correspondence follows from the fact that $d_H(X, Y) < r$. The desired estimate $\operatorname{dis}\mathfrak{R} < 2r$ follows from the triangle inequality: if $(x, y) \in \mathfrak{R}$ and $(x', y') \in \mathfrak{R}$, then

$$|d(x, x') - d(y, y')| \le d(x, y) + d(x', y') < 2r.$$

2. $d_{GH}(X, Y) \le \frac{1}{2} \operatorname{dis}\mathfrak{R}$ for any correspondence $\mathfrak{R}$. Let $\operatorname{dis}\mathfrak{R} = 2r$. To avoid confusion, we use the notation d_X and d_Y for the metrics of X and Y,

respectively. It suffices to show that there is a semi-metric d on the disjoint union $X \cup Y$ such that $d|_{X\times X} = d_X$, $d|_{Y\times Y} = d_Y$, and $d_H(X,Y) \le r$ in $(X \cup Y, d)$. The idea is to set the distance from x to y equal to r whenever x and y correspond to each other, and then take the minimal metric d generated by this condition. This is formally achieved as follows: for an $x \in X$ and $y \in Y$ define

$$d(x,y) = \inf\{d_X(x,x') + r + d_Y(y',y) : (x',y') \in \mathfrak{R}\}$$

(the distances within X and Y are already defined by d_X and d_Y). Verifying the triangle inequality for d and the fact that $d_H(X,Y) \le r$ is left as an exercise to the reader. □

Exercise 7.3.26. Let X, Y and Z be metric spaces, $\mathfrak{R}_1$ be a correspondence between X and Y, and $\mathfrak{R}_2$ be a correspondence between Y and Z. A *composition* of $\mathfrak{R}_1$ and $\mathfrak{R}_2$, denoted by $\mathfrak{R}_1 \circ \mathfrak{R}_2$, is set of all $(x,z) \in X \times Z$ for which there is a $y \in Y$ such that $(x,y) \in \mathfrak{R}_1$ and $(y,z) \in \mathfrak{R}_2$.

1. Prove that $\mathfrak{R}_1 \circ \mathfrak{R}_2$ is indeed a correspondence.
2. Prove that $\operatorname{dis}(\mathfrak{R}_1 \circ \mathfrak{R}_2) \le \operatorname{dis}\mathfrak{R}_1 + \operatorname{dis}\mathfrak{R}_2$.
3. Using the above inequality, give an alternative proof of the triangle inequality for the Gromov–Hausdorff distance.

The next corollary gives us more techniques for handling the Gromov–Hausdorff distances. While it does not give explicit expressions for the distance, it provides another quantity which differs from the distance in no more than two times. Note that an estimate of this type is sufficient to study the *topology* (on the space of metric spaces) determined by the Gromov–Hausdorff distance.

Definition 7.3.27. Let X and Y be metric spaces and $\varepsilon > 0$. A (possibly noncontinuous!) map $f\colon X \to Y$ is called an ε-isometry if $\operatorname{dis} f \le \varepsilon$ and $f(X)$ is an ε-net in Y.

Corollary 7.3.28. *Let X and Y be two metric spaces and $\varepsilon > 0$. Then*

1. *If $d_{GH}(X,Y) < \varepsilon$, then there exists a 2ε-isometry from X to Y.*
2. *If there exists an ε-isometry from X to Y, then $d_{GH}(X,Y) < 2\varepsilon$.*

Proof. 1. Let $\mathfrak{R}$ be a correspondence between X and Y with $\operatorname{dis}\mathfrak{R} < 2\varepsilon$. For every $x \in X$, choose $f(x) \in Y$ such that $(x, f(x)) \in \mathfrak{R}$. This defines a map $f\colon X \to Y$. Obviously $\operatorname{dil} f \le \operatorname{dil}\mathfrak{R} < 2\varepsilon$. Let us show that $f(X)$ is a 2ε-net in Y. For a $y \in Y$, consider an $x \in X$ such that $(x,y) \in \mathfrak{R}$. Since both y and $f(x)$ are in correspondence with x, one has $d(y, f(x)) \le d(x,x) + \operatorname{dis}\mathfrak{R} < 2r$. Hence $\operatorname{dist}(y, f(X)) < 2r$.

2. Let f be an ε-isometry. Define $\mathfrak{R} \subset X \times Y$ by

$$\mathfrak{R} = \{(x, y) \in X \times Y : d(y, f(x)) \leq \varepsilon\}.$$

Then $\mathfrak{R}$ is a correspondence because $f(X)$ is an ε-net in Y. If $(x, y) \in \mathfrak{R}$ and $(x', y') \in \mathfrak{R}$, one has

$$\begin{aligned}|d(y, y') - d(x, x')| &\leq |d(f(x), f(x')) - d(x, x')| + d(y, f(x)) + d(y', f(x')) \\ &\leq \operatorname{dis} f + \varepsilon + \varepsilon \leq 3\varepsilon.\end{aligned}$$

Hence $\operatorname{dis} \mathfrak{R} \leq 3r$, and Theorem 7.3.25 implies $d_{GH}(X, Y) \leq \frac{3}{2} r < 2r$. □

Remark 7.3.29. It is important that we do not require continuity of ε-isometries. Even if two spaces are very close with respect to the Gromov–Hausdorff distance, it can happen that there are no continuous maps with small distortion—recall the spheres with small handles in Subsection 7.1.4.

Now we are in a position to prove the analog of Theorem 7.2.4 for the Gromov–Hausdorff distance. Note that, unlike Lipschitz distance, the Gromov–Hausdorff one defines a *finite* metric.

Theorem 7.3.30. *Gromov–Hausdorff distance defines a finite metric on the space of isometry classes of compact metric spaces. In other words, it is nonnegative, symmetric and satisfies the triangle inequality; moreover* $d_{GH}(X, Y) = 0$ *if and only if* X *and* Y *are isometric.*

Proof. We have already proven all statements of this theorem except the claim that $d_{GH}(X, Y) = 0$ implies that X and Y are isometric. Let X and Y be two compact spaces such that $d_{GH}(X, Y) = 0$. By Corollary 7.3.28, there exists a sequence of maps $f_n : X \to Y$ such that $\operatorname{dis} f_n \to 0$. Fix a countable dense set $S \subset X$. Using the Cantor diagonal procedure, one can choose a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ such that for every $x \in S$ the sequence $\{f_{n_k}(x)\}$ converges in Y. Without loss of generality we may assume that this holds for $\{f_n\}$ itself. Then one can define a map $f : S \to Y$ as the limit of f_n, namely, set $f(x) = \lim f_n(x)$ for every $x \in S$. Since $|d(f_n(x), f_n(y)) - d(x, y)| \leq \operatorname{dis} f_n \to 0$, we have $d(f(x), f(y)) = \lim d(f_n(x), f_n(y)) = d(x, y)$ for all $x, y \in S$. In other words, f is a distance-preserving map from S to Y. Then f can be extended to a distance-preserving map from the entire X to Y (by Proposition 1.5.9). Now we can finish the proof in the same way as in Theorem 7.2.4. Namely there is a similar distance-preserving map from Y to X, and it follows that X and Y are isometric. □

The above theorem allows us to consider compact metric spaces as points in the so-called *Gromov–Hausdorff space*, keeping in mind that isometric spaces represent the same "point". The topology of this space (determined

by the Gromov–Hausdorff distance) is called the *Gromov–Hausdorff topology*. (The set-theoretic remarks that we made after Theorem 7.2.4 apply here as well.)

Exercise 7.3.31. Prove the following generalization of Theorem 7.3.30: if X and Y are metric spaces with $d_{GH}(X,Y) = 0$, X is compact and Y is complete, then X and Y are isometric.

Hint: Show that Y is compact. To do that, construct a finite ε-net in Y for every $\varepsilon > 0$.

7.4. Gromov–Hausdorff Convergence

In this section we consider converging sequences in the Gromov–Hausdorff space of compact metric spaces. By definition, a sequence $\{X_n\}_{n=1}^{\infty}$ of compact metric spaces converges to a compact metric space X if $d_{GH}(X_n, X) \to 0$ as $n \to \infty$. In this case, we will write $X_n \xrightarrow[GH]{} X$ and call X the *Gromov–Hausdorff limit* of $\{X_n\}$. Since d_{GH} is a metric (Theorem 7.3.30), the limit is unique up to an isometry.

7.4.1. Examples and properties.

Example 7.4.1 (Hausdorff convergence)**.** For subspaces in the same metric space, Gromov–Hausdorff distance by definition is not greater than the Hausdorff distance. Thus Hausdorff convergence of subsets of a given space implies Gromov–Hausdorff convergence (but not vice versa).

Corollary 7.3.28 yields the following criterion for Gromov–Hausdorff convergence: a sequence $\{X_n\}$ of metric spaces converges to a metric space X if and only if there are a sequence $\{\varepsilon_n\}$ of numbers and a sequence of maps $f_n \colon X_n \to X$ (or, alternatively, $f_n \colon X \to X_n$) such that every f_n is an ε_n-isometry and $\varepsilon_n \to 0$. This observation includes "functional" types of convergence in our list of examples.

Example 7.4.2 (uniform convergence)**.** If a sequence $\{X_n\}$ of metric spaces uniformly converges (in the sense of Definition 7.1.5) to a metric space X, then $X_n \xrightarrow[GH]{} X$.

Example 7.4.3 (Lipschitz convergence)**.** As we have seen in Exercise 7.2.6, the Lipschitz convergence is a particular case of uniform convergence, and hence a particular case of the Gromov–Hausdorff convergence. In other words, the Gromov–Hausdorff topology is weaker than the Lipschitz one.

Example 7.4.4 (uniform convergence to a semi-metric)**.** Let $\{d_n\}$ be a sequence of metrics on a fixed set X which converges uniformly to some function $d \colon X \times X \to \mathbb{R}$. Then d is obviously a semi-metric, and the quotient

metric space X/d (cf. Proposition 1.1.5) is the Gromov–Hausdorff limit of the spaces (X, d_n).

Indeed, the distortions of the projection $X \to X/d$ with respect to the metrics d_n go to zero.

Note that if X is a finite set, it suffices to require that $d_n(x, y) \to d(x, y)$ for every pair $x, y \in X$ because a pointwise convergence of functions on a finite set implies uniform convergence.

Exercise 7.4.5. Prove that any converging sequence $\{X_n\}$ of finite metric spaces of a fixed cardinality N can be represented in the form given in the above example. In other words, there is a sequence of metrics $\{d_n\}$ on a fixed set X such that (X, d_n) is isometric to X_n and $\{d_n\}$ uniformly converges to a semi-metric.

In particular, the limit of $\{X_n\}$ is a finite space of cardinality no greater than N.

Exercise 7.4.6. Prove that a sequence $\{X_n\}$ converges to a one-point space if and only if $\operatorname{diam}(X_n) \to 0$.

The next exercise gives a general form of a sequence converging to a finite space.

Exercise 7.4.7. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of metric spaces, and let X be a finite metric space of cardinality N, $X = \{x_i : 1 \le i \le N\}$.

1. Assume that $X_n \xrightarrow[GH]{} X$. Prove that, for all sufficiently large n, the cardinality of X_n is at least N.

2. Prove that $X_n \xrightarrow[GH]{} X$ if and only if the following holds. For all sufficiently large n, X_n can be split into a disjoint union of N nonempty sets $X_{n,1}, X_{n,2}, \ldots, X_{n,N}$ so that for all i, j

$$\operatorname{diam}(X_{n,i}) \to 0, \quad \operatorname{dist}(X_{n,i}, X_{n,j}) \to |x_i x_j| \qquad (n \to \infty).$$

Exercise 7.4.8. Let N be a fixed natural number. Prove that the Lipschitz, uniform, and Gromov–Hausdorff distances determine the same topology on the class of finite metric spaces of cardinality N. Give an explicit description of this topology.

The reader may wonder why we pay attention to the trivial case of finite spaces. One of the reasons is that finite spaces form a dense set in the Gromov–Hausdorff space:

Example 7.4.9. Every compact metric space X is a limit of finite spaces. Indeed, take a sequence $\varepsilon_n \to 0$ of positive numbers and choose a finite ε_n-net S_n in X for every n. Then $S_n \xrightarrow[GH]{} X$, simply because $d_{GH}(X, S_n) \le d_H(X, S_n) \le \varepsilon_n$.

Moreover, taking appropriate ε-nets one can essentially reduce convergence of arbitrary compact metric spaces to convergence of their finite subsets. The details follow.

Definition 7.4.10. Let X and Y be two compact metric spaces, and $\varepsilon, \delta > 0$ We say that X and Y are (ε, δ)-approximations of each other if there exist finite collections of points $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ in X and Y, respectively, such that:

(1) The set $\{x_i : 1 \le i \le N\}$ is an ε-net in X, and $\{y_i : 1 \le i \le N\}$ is an ε-net in Y.

(2) $|d_X(x_i, x_j) - d_Y(y_i, y_j)| < \delta$ for all $i, j \in \{1, \dots, N\}$.

An *ε-approximation* is a synonym for $(\varepsilon, \varepsilon)$-approximation (i.e., we omit the δ if $\delta = \varepsilon$).

Proposition 7.4.11. *Let X and Y be compact metric spaces.*

(1) *If Y is an (ε, δ)-approximation of X, then $d_{GH}(X, Y) < 2\varepsilon + \delta$.*

(2) *If $d_{GH}(X, Y) < \varepsilon$, then Y is a 5ε-approximation of X.*

Proof. 1. Let $X_0 = \{x_i\}_{i=1}^N$ and $Y_0 = \{y_i\}_{i=1}^N$ be as in Definition 7.4.10. The second condition in the definition means that the natural correspondence $\{(x_i, y_i) : 1 \le i \le N\}$ between X_0 and Y_0 has distortion less than δ. It follows that $d_{GH}(X_0, Y_0) < \delta/2$. Since X_0 and Y_0 are ε-nets in X and Y, respectively, we have $d_{GH}(X, X_0) \le \varepsilon$ and $d_{GH}(Y, Y_0) \le \varepsilon$. The statement follows by the triangle inequality for d_{GH}.

2. By Corollary 7.3.28, there is a 2ε-isometry $f\colon X \to Y$. Let $X_0 = \{x_i\}_{i=1}^N$ be an ε-net in X and $y_i = f(x_i)$. Then $|d(x_i, x_j) - d(y_i, y_j)| < 2\varepsilon < 5\varepsilon$ for all i, j. It remains to prove that $Y_0 = \{y_i : 1 \le i \le N\}$ is a 5ε-net in Y. Pick a $y \in Y$. Since $f(X)$ is a 2ε-net in Y, there is an $x \in X$ such that $d(y, f(x)) \le 2\varepsilon$. Since X_0 is an ε-net in X, there exists an $x_i \in X_0$ such that $d(x, x_i) \le \varepsilon$. Then

$$\begin{aligned} d(y, y_i) = d(y, f(x_i)) &\le d(y, f(x)) + d(f(x), f(x_i)) \\ &\le 2\varepsilon + d(x, x_i) + \operatorname{dis} f \le 2\varepsilon + \varepsilon + 2\varepsilon \le 5\varepsilon. \end{aligned}$$

Hence $\operatorname{dist}(y, Y_0) \le d(y, y_i) \le 5\varepsilon$. □

The above proposition yields a criterion for convergence: $X_n \xrightarrow[GH]{} X$ if and only if, for any $\varepsilon > 0$, X_n is an ε-approximation of X for all large enough n. There is a more elegant formulation of this kind:

Proposition 7.4.12. *For compact metric spaces X and $\{X_n\}_{n=1}^\infty$, $X_n \xrightarrow[GH]{}$ X if and only if the following holds. For every $\varepsilon > 0$ there exist a finite ε-net S in X and an ε-net S_n in each X_n such that $S_n \xrightarrow[GH]{} S$.*

Moreover these ε-nets can be chosen so that, for all sufficiently large n, S_n have the same cardinality as S.

Proof. If such ε-nets exist, then X_n is an ε-approximation of X for all sufficiently large n. Then $X_n \xrightarrow[GH]{} X$ by the previous proposition. To prove the converse implication, take a finite $(\varepsilon/2)$-net S in X and construct corresponding nets S_n in X_n. Namely, pick a sequence of ε_n-approximations $f_n\colon X \to X_n$ where $\varepsilon_n \to 0$ and define $S_n = f_n(S)$. Then $S_n \xrightarrow[GH]{} S$ and, as in the previous proposition, S_n is an ε-net in X_n for all large enough n. $\square$

Let us mention one important implication of this construction. Let $X_n \xrightarrow[GH]{} X$ and S be a finite subset of X (not necessary an ε-net for a small ε). As in the above proof, construct sets $S_n \subset X_n$ corresponding to S. Then S_n converge to X; i.e., the distances in S_n converge to the corresponding distances in S. It follows that all (reasonable) geometric characteristics of the sets S_n converge to those of S. This opens a way to prove various continuity statements about the Gromov–Hausdorff space: if some property of spaces X_n can be formulated in terms of finite collections of points, then this property is automatically inherited by the limit space X.

To see how this abstract scheme works, consider the property of a metric to be intrinsic. We have a midpoint criterion (Theorem 2.4.16) expressing this property in terms of triples of points. Since triples of points in converging spaces correspond to almost isometric triples in the limit space, it follows that a limit of compact length spaces is a length space. We will repeat this proof in more formal words in Section 7.5 (see Theorem 7.5.1).

7.4.2. Compactness theorem. Since the Gromov–Hausdorff topology is a relatively weak one (compared to, say, the topology of the Lipschitz distance), one may expect that it has relatively many compact sets. Indeed, many natural classes of metric spaces form (pre-)compact sets in the Gromov-Hausdorff space. In this section we do not prove any statements of this sort, but give a general criterion for pre-compactness.

Proposition 7.4.12 implies that members of a sequence $\{X_n\}$ converging in the Gromov–Hausdorff space must contain ε-nets of uniformly bounded cardinality (for every given $\varepsilon > 0$). It follows that, if a class $\mathfrak{X}$ of metric spaces is pre-compact in the Gromov–Hausdorff topology, then for every $\varepsilon > 0$ the size of a minimal ε-net is uniformly bounded over all elements of $\mathfrak{X}$. It turns out that this property of $\mathfrak{X}$, along with the fact that the diameters of its members are uniformly bounded, is sufficient for pre-compactness.

Definition 7.4.13. We say that a class $\mathfrak{X}$ of compact metric spaces is *uniformly totally bounded* if

(1) There is a constant D such that $\operatorname{diam} X \le D$ for all $X \in \mathfrak{X}$.

(2) For every $\varepsilon > 0$ there exists a natural number $N = N(\varepsilon)$ such that every $X \in \mathfrak{X}$ contains an ε-net consisting of no more than N points.

Exercise 7.4.14. Prove that the first condition of the above definition is redundant (i.e., is implied by the second one) if all elements of $\mathfrak{X}$ are length spaces.

Theorem 7.4.15. *Any uniformly totally bounded class $\mathfrak{X}$ of compact metric spaces is pre-compact in the Gromov–Hausdorff topology. That is, any sequence of elements of $\mathfrak{X}$ contains a converging subsequence.*

Proof. Let D and $N(\varepsilon)$ be as in Definition 7.4.13. Define $N_1 = N(1)$ and $N_k = N_{k-1} + N(1/k)$ for all $k \geq 2$. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of metric spaces from $\mathfrak{X}$. In every space X_n, consider a union of $(1/k)$-nets for all $k \in \mathbb{N}$. This is a countable dense collection $S_n = \{x_{i,n}\}_{i=1}^{\infty} \subset X_n$ such that for every k the first N_k points of S_n form a $(1/k)$-net in X_n. The distances $|x_{i,n}x_{j,n}|$ do not exceed D, i.e., belong to a compact interval. Therefore, using the Cantor diagonal procedure, we can extract a subsequence of $\{X_n\}$ in which $\{|x_{i,n}x_{j,n}|\}_{n=1}^{\infty}$ converge for all i, j. To simplify the notation, we assume that they converge without passing to a subsequence.

Now let us construct the limit space $\bar{X}$ for $\{X_n\}$. Pick an abstract countable set $X = \{x_i\}_{i=1}^{\infty}$ and define a semi-metric d on X by

$$d(x_i, x_j) = \lim_{n\to\infty} |x_{n,i}x_{n,j}|.$$

The quotient construction from Proposition 1.1.5 gives us a metric space X/d. We will denote by $\bar{x}_i$ the point of X/d obtained from x_i. This quotient space may not be complete, so let $\bar{X}$ be the completion of X/d. We will prove that $\{X_n\}$ converges to $\bar{X}$.

For a $k \in N$, consider the set $S^{(k)} = \{\bar{x}_i : 1 \leq i \leq N_k\} \subset \bar{X}$. It is a $(1/k)$-net in $\bar{X}$. Indeed, every set $S_n^{(k)} = \{x_{i,n} : 1 \leq i \leq N_k\}$ is a $(1/k)$-net in the respective space X_n. Hence for every $x_{i,n} \in S_n$ there is a $j \leq N_k$ such that $|x_{i,n}x_{j,n}| \leq 1/k$. Since N_k do not depend on n, for every fixed $i \in N$ there is a $j \leq N_k$ such that $|x_{i,n}x_{j,n}| \leq 1/k$ for infinitely many indices n. Passing to the limit we obtain that $|\bar{x}_i\bar{x}_j| \leq 1/k$ for this j. Thus $S^{(n)}$ is a $(1/k)$-net in X/d and hence in $\bar{X}$. Since $\bar{X}$ is complete and has a $(1/k)$-net for any $k \in N$, it is compact.

Furthermore, the set $S^{(k)}$ is a Gromov–Hausdorff limit of the sets $S_n^{(k)}$ as $n \to \infty$, because these are finite sets consisting of N_k points (some of which may coincide) and the distances converge. Thus for every $k \in \mathbb{N}$ we have a $(1/k)$-net in $\bar{X}$ which is a Gromov–Hausdorff limit of some $(1/k)$-nets in the spaces X_n. By Proposition 7.4.12 it follows that $X_n \underset{GH}{\longrightarrow} \bar{X}$. □

Exercise 7.4.16. Let M^n be a compact Riemannian manifold. Prove that for any sufficiently small $\varepsilon > 0$ there is an ε-net in M containing no more than $\varepsilon^{-n}C(n)\,\mathrm{Vol}(M)$ points (for some constant $C(n)$ independent of M). Here Vol denotes the Riemannian volume.

The result of the above exercise does *not* mean that any class of compact Riemannian n-manifolds with uniformly bounded volumes is pre-compact in the Gromov–Hausdorff topology. The catch is in the phrase "sufficiently small"— the actual bound for being "small" depends on M.

Exercise 7.4.17. Give an example of a set of two-dimensional compact Riemannian manifolds with areas no greater than 1 which is not pre-compact in the Gromov–Hausdorff topology.

Information. Here are some important examples (without proofs) of pre-compact classes of Riemannian manifolds.

Bounded volume and injectivity radius. For any $n \in \mathbb{N}$ and any $r, V > 0$, the class of all n-dimensional Riemannian manifolds with volume $\le V$ and injectivity radius $\ge r$, is pre-compact.

Bounded diameter and curvature. For any $n \in N$ and any $\kappa \in \mathbb{R}$, $D > 0$, the class of all n-dimensional Riemannian manifolds with diameter $\le D$ and sectional curvature $\ge \kappa$ is pre-compact. We will prove this in Chapter 10 as a part of a more general statement about Alexandrov spaces (Theorem 10.7.2). Moreover, the same is true with Ricci curvature instead of sectional one.

7.5. Convergence of Length Spaces

We already mentioned that a Gromov–Hausdorff limit of length spaces is a length space. In other words, the length spaces form a closed set in the Gromov–Hausdorff topology. Here we state this important fact as a theorem and give a more detailed proof (which is nothing but a formalization of the argument given at the end of Subsection 7.4.1).

Theorem 7.5.1. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of length spaces, X a complete metric space, and $X_n \xrightarrow[GH]{} X$. Then X is a length space.*

Proof. By the criterion for a complete length metric (Theorem 2.4.16), it suffices to prove that any two points $x, y \in X$ possess an ε-midpoint for any $\varepsilon > 0$. Let n be such that $d_{GH}(X, X_n) < \varepsilon/10$. Then, by Theorem 7.3.25, there is a correspondence $\mathfrak{R}$ between X and X_n whose distortion is less than $\varepsilon/5$. Take points $\tilde{x}, \tilde{y} \in X_n$ corresponding to x and y. Since X_n is a length space, there is a $\tilde{z} \in X_n$ which is an $(\varepsilon/5)$-midpoint for $\tilde{x}$ and $\tilde{y}$. Let $z \in X$

be a point corresponding to $\tilde{z}$. Then

$$\big||xz| - \tfrac{1}{2}|xy|\big| \le \big||\tilde{x}\tilde{z}| - \tfrac{1}{2}|\tilde{x}\tilde{y}|\big| + 2\,\mathrm{dis}\,\mathfrak{R} < \varepsilon/5 + 2\varepsilon/5 < \varepsilon.$$

Similarly $\big||yz| - \frac{1}{2}|xy|\big| < \varepsilon$. Thus z is an ε-midpoint for x and y. □

7.5.1. First examples. One should be careful when verifying convergence of length spaces, because distances in length metrics is a complicated subject. This warning message is explained by the following exercises, which should help the reader to understand what kind of difficulties can arise.

Exercise 7.5.2. 1. Let X_n be the sphere S^2 with a ball of radius $1/n$ removed. Prove that the spaces X_n (regarded with their intrinsic metrics) converge to S^2.

2. Let X_n be obtained the same way from the circle S^1. Show that X_n do *not* converge to S^1.

Exercise 7.5.3. 1. Let X be a straight line segment in $\mathbb{R}^3$, and let X_n be the boundary of its $(1/n)$-neighborhood (it is a two-dimensional surface), equipped with the length metric induced from $\mathbb{R}^3$. Prove that $X_n \xrightarrow[GH]{} X$ as $n \to \infty$.

2. Let X be a planar disc in $\mathbb{R}^3$, and again let X_n be the boundary of its $(1/n)$-neighborhood equipped with the induced length metric. Prove that the sequence $\{X_n\}_{n=1}^{\infty}$ converges in the Gromov–Hausdorff sense, but the limit is *not* isometric to X.

Exercise 7.5.4. Justify the convergence of "spheres with small handles" described in Subsection 7.1.4. Namely, let X_n be a length space obtained from the standard unit sphere by removing a round disc of diameter less than $1/n$ and attaching a handle whose (intrinsic!) diameter is less than $1/n$. Prove that $X_n \xrightarrow[GH]{} S^2$.

While any compact metric space can be obtained as a limit of finite spaces (Example 7.4.9), these finite spaces do not carry length metrics. For length spaces, the role of finite (zero-dimensional) spaces is played by the one-dimensional ones, i.e., graphs. Recall that a finite metric graph is a length space obtained by gluing together several spaces isometric to line segments in such a way that only endpoints may be shared between the segments. Equivalently, a finite metric graph is a finite topological graph equipped with a length metric.

Proposition 7.5.5. *Every compact length space can be obtained as a Gromov–Hausdorff limit of finite graphs.*

Proof. Let X be the length space in question. Pick small positive numbers ε and δ (where δ is much smaller than ε), and choose a finite δ-net S in X.

Then consider the following graph G: the set of vertices of G is S, two points $x, y \in S$ are connected by an edge if and only if $|xy| < \varepsilon$, and the length of this edge equals $|xy|$.

Let us show that G is an ε-approximation for X (cf. Definition 7.4.10) if δ is small enough, say, $\delta < \frac{1}{4}\varepsilon^2/\mathrm{diam}(X)$. We consider S both as a subset of X and a subset of G. Obviously S is an ε-net in both spaces, and $|xy|_G \geq |xy|$ for all $x, y \in S$ where $|\cdot|_G$ denotes the distance in G. It remains to show that $|xy|_G \leq |xy| + \varepsilon$.

Let γ be a shortest path in X connecting x and y. Choose n points $x_1, \dots, x_n$, where $n \leq 2L(\gamma)/\varepsilon$, dividing γ into intervals of lengths no greater than $\varepsilon/2$. For every $i = 1, \dots, n$, there is a point $y_i \in S$ such that $|x_i y_i| \leq \delta$. In addition, set $x_0 = y_0 = x$ and $x_{n+1} = y_{n+1} = y$. Note that $|y_i y_{i+1}| \leq |x_i x_{i+1}| + 2\delta < \varepsilon$ for all $i = 0, \dots, n$. In particular, so y_i and y_{i+1} are connected by an edge in G provided that $\delta < \varepsilon/4$. Then

$$|xy|_G \leq \sum_{i=0}^{n} |y_i y_{i+1}| \leq \sum_{i=0}^{n} |y_i y_{i+1}| + 2\delta n = |xy| + 2\delta n.$$

Recall that $n \leq 2L(\gamma)/\varepsilon \leq 2\,\mathrm{diam}(X)/\varepsilon$; hence

$$|xy|_G \leq |xy| + \delta \cdot \frac{4\,\mathrm{diam}(X)}{\varepsilon} < |xy| + \varepsilon$$

if $\delta < \frac{1}{4}\varepsilon^2/\mathrm{diam}(X)$.

Thus we have a finite graph which is an ε-approximation for X. Passing ε to zero yields a sequence of graphs converging to X. □

Exercise 7.5.6. Prove that every compact length space X can be represented as a Gromov–Hausdorff limit of finite graphs isometrically embedded in X, i.e., of topological graphs embedded in X and equipped with their induced length metrics.

Hint: Utilize the same construction as in the above proof, but draw the edges in X (as shortest paths) adding new vertices when these shortest paths intersect one another. To get rid of weird cases when one has to add infinitely many vertices, show that shortest paths can be chosen so that the intersection of any two of them is the empty set, or a single point, or an interval of both paths.

Note that the number of vertices and edges of graphs constructed in the proof of Proposition 7.5.5 tends to infinity as ε approaches zero. In other words, the graphs get more and more complex. This is not a defect of this particular construction but a consequence of a general obstacle:

Exercise 7.5.7. 1. Let N be a natural number and $\{X_n\}_{n=1}^{\infty}$ a sequence of graphs each having no more than N edges. Prove that the limit of $\{X_n\}$, if one exists, is a finite graph.

2. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of finite graphs each having no more than N vertices. Prove that the limit of $\{X_n\}$, if one exists, is a (possibly infinite) graph.

7.5.2. Topology of Gromov–Hausdorff limits. This subsection consists of a series of exercises, some of which are quite complicated. They give some examples of what can and what cannot happen to the Gromov–Hausdorff limit of length spaces.

A general rule of thumb: in dimensions 1 and 2, one can expect nice topological relations between converging spaces and their limit. In higher dimensions, there are relations between fundamental groups but not much can be said beyond this.

The first exercise is a general fact about Gromov–Hausdorff convergence (not specific to length spaces). To get better understanding of it, recall again the spheres with vanishing handles from Section 7.1.4. There are no continuous maps with small distortion from the sphere to a sphere with a small handle, but such a map in the opposite direction exists: just project the surface onto the sphere.

Exercise 7.5.8. (a) Let X be the unit ball in $\mathbb{R}^n$. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of compact metric spaces, $X_n \xrightarrow[GH]{} X$. Prove that there exists a sequence of *continuous* maps $f_n : X_n \to X$ such that each f_n is an ε_n-isometry for some sequence $\varepsilon_n \to 0$.

Prove the same if X is

(b) the unit circle S^1;

(c) the sphere S^{n-1};

(d) the n-dimensional torus $T^n = S^1 \times \cdots \times S^1$;

(e) a metric space homeomorphic to one of the above;

(f) a metric space homeomorphic to a compact smooth manifold.

(g) Prove the same under some weaker (as weak as possible) topological restrictions on X.

Hint: Use Corollary 7.3.28 to obtain a sequence of possibly discontinuous maps, then approximate these maps by continuous ones.

Exercise 7.5.9. Let $\{X_n\}$ be a sequence of compact length spaces, $X_n \xrightarrow[GH]{} S^1$. Prove that, for all large enough n, the spaces X_n are not simply connected.

Hint: Let $\{f_n\}$ be a sequence of maps from the previous exercise. Construct an "approximate lift" of the circle into X_n, namely, a closed curve in X_n that is mapped onto the circle "almost isometrically" in some sense. Then show that this curve is not contractible.

Remark 7.5.10. It is essential that X_n in the above exercise are length spaces. To see why, consider a sequence $\{X_n\}$ where X_n is an arc in S^1 of length $2\pi - \frac{1}{n}$ equipped with the (nonintrinsic) restriction of the circle's metric.

Exercise 7.5.11. Let $\{X_n\}$ be a sequence of compact length spaces, $X_n \xrightarrow[GH]{} X$, and let f_n be as in Exercise 7.5.9. Prove that, for all large enough n, f_n induces a surjective homomorphism of fundamental groups.

In particular, if the spaces X_n are simply connected, then X is simply connected too.

Hint: See the previous exercise.

Exercise 7.5.12. Let B be the unit two-dimensional disk with its standard Euclidean metric. Let $\{B_n\}$ be a sequence of length spaces homeomorphic to B, $B_n \xrightarrow[GH]{} B$, and let ε be a positive number. Prove that for all large enough n, there is a point $p_n \in B_n$ such that $\operatorname{dist}(p_n, \partial B_n) > 1 - \varepsilon$.

Hint: Prove that there is a Jordan curve in B_n whose image in B is close to ∂B (in the sense of uniform distance). This curve bounds a region in B_n. Within that region, there must be a point corresponding to the center of B in the sense of Gromov–Hausdorff approximation. This point can be taken for p_n.

Exercise 7.5.13. Prove that a sequence of length spaces homeomorphic to the sphere S^2 cannot converge to (a) the standard two-dimensional disc B; (b) a space homeomorphic to B.

Exercise 7.5.14. Let $\{X_n\}$ be a sequence of length spaces homeomorphic to S^2, $X_n \xrightarrow[GH]{} X$, and X is also homeomorphic to S^2. Prove that there is a sequence of homeomorphisms $f_n : X_n \to X$ with $\operatorname{dis}(f_n) \to 0$.

Exercise 7.5.15. 1. Let $\{X_n\}$ be a sequence of length spaces homeomorphic to S^2 converging to some compact space X. Prove that X cannot contain a subset homeomorphic to the three-dimensional disc.

2. Prove the same for a sequence $\{X_n\}$ of compact two-dimensional manifolds with length metrics, provided that the genus of X_n (the number of handles or films) is uniformly bounded.

3. Show that, without the condition of uniformly bounded genus, the limit of $\{X_n\}$ can be any compact length space. (Compare this with Proposition 7.5.5 and Exercise 7.5.7.)

Hint to 1 and 2: There exists a graph that can be topologically embedded into a three-dimensional disc (any graph can), but not into X_n.

Most of the properties given in the above group of exercises are only valid in the two-dimensional case. Below are some counterexamples to their higher-dimensional counterparts.

Exercise 7.5.16. Let B be the standard three-dimensional ball. Show that, for any given $\varepsilon > 0$, there exists a length space B' homeomorphic to B such that $d_{GH}(B, B') < \varepsilon$ and $\partial B'$ is an ε-net in B'. (Compare with Exercise 7.5.12.)

Hint: Push the boundary inside the ball so that it becomes an ε-net but the intrinsic distances in the ball do not change much.

Exercise 7.5.17. Construct a sequence $\{X_n\}$ of length spaces homeomorphic to S^3 and converging to the three-dimensional ball. (Compare with Exercise 7.5.13.)

Hint: Consider doublings of the spaces B' from the previous exercise (i.e., the result of gluing together two isometric copies of B' along the boundary.)

Exercise 7.5.18. Construct a sequence $\{X_n\}$ of length spaces homeomorphic to S^3 such that $X_n \xrightarrow[GH]{} S^3$ but no homeomorphism from X_n to S^3 has distortion less than $1/10$. (Compare with Exercise 7.5.15.)

Chapter 8

Large-scale Geometry

To get an idea of what we are going to talk about in this chapter, imagine a device measuring distances with a precision, say, one mile. Such a tool is useless for an engineer investigating the shape of a car, but it is more than excellent for learning geometry of the solar system. In this chapter we consider metric properties for which such a measuring device is perfectly good. In other words, we are not going to distinguish two metrics if the difference between them is bounded by a constant. More precisely, the properties that we will discuss are the same for spaces lying within finite Gromov–Hausdorff distance from one another. For example, a Euclidean space and a lattice in it (regarded with the restriction of the Euclidean metric) look the same from this point of view. Of course, no local properties survive through such a transformation of a space but many global and asymptotic ones remain.

In some cases, we will admit even less precise "measurement of distances". For example, consider a measuring instrument which may give the result ten times greater or smaller than the actual distance, plus the same one-mile error. Quite surprisingly, this instrument allows one, for example, to tell the Euclidean plane from the hyperbolic one. (A question of this sort, concerning the physical universe, is a famous problem in modern cosmogony and physics. Alas, our real instruments are not that good.)

8.1. Noncompact Gromov–Hausdorff Limits

For noncompact spaces, convergence with respect to the Gromov–Hausdorff distance is not very useful. There is an analogy with the uniform convergence of functions on a fixed domain. While the domain is compact, uniform

convergence is a powerful and widely used notion, but it becomes too restrictive once noncompact domains come into question. For example, a sequence $\{\lambda_n f\}$, where f is a continuous function and $\{\lambda_n\}$ is a converging sequence of real numbers, may fail to converge uniformly due to the fact that f may not be bounded. Instead, one should utilize the wider notion of uniform convergence on compact sets: a sequence of functions converges if it converges uniformly on every compact subset of the domain.

A similar approach is taken to the Gromov–Hausdorff convergence when noncompact metric spaces are involved. Roughly speaking, a sequence $\{X_n\}$ of metric spaces converges to a space X if for every $r > 0$ the balls of radius r in X_n centered at some fixed points converge (as compact metric spaces) to a ball of radius r in X. The actual definition (Definition 8.1.1 below) is more complicated, but in most cases it is equivalent to this description.

To see the difference from the ordinary Gromov–Hausdorff convergence, consider a sequence of Euclidean spheres of radii growing to infinity. For every given $r > 0$, the sets of diameter no greater than r in these spheres look more and more similar to subsets of the Euclidean plane, but the whole spheres do not get close to the plane in any sense. These spheres have no limit with respect to the Gromov–Hausdorff distance, but they converge to the plane in the extended sense that we are about to define.

Definition 8.1.1. A *pointed metric space* is a pair (X, p) consisting of a metric space X and a point $p \in X$.

A sequence $\{(X_n, p_n)\}_{n=1}^{\infty}$ of pointed metric spaces converges in the Gromov–Hausdorff sense to a pointed metric space (X, p) if the following holds. For every $r > 0$ and $\varepsilon > 0$ there exists a natural n_0 such that for every $n > n_0$ there is a (not necessarily continuous) map $f : B_r(p_n) \to X$ such that the following hold:

(1) $f(p_n) = p$;

(2) $\operatorname{dis} f < \varepsilon$;

(3) the ε-neighborhood of the set $f(B_r(p_n))$ contains the ball $B_{r-\varepsilon}(p)$.

We will use the notation $(X_n, p_n) \xrightarrow[GH]{} (X, p)$ for this type of convergence.

Exercise 8.1.2. Prove that for compact metric spaces the convergence of pointed spaces is equivalent to the ordinary Gromov–Hausdorff convergence in the following sense:

1. $(X_n, p_n) \xrightarrow[GH]{} (X, p)$ implies that $X_n \xrightarrow[GH]{} X$.

2. If $X_n \xrightarrow[GH]{} X$ and $p \in X$, then one can choose a point p_n in every X_n so that $(X_n, p_n) \xrightarrow[GH]{} (X, p)$.

The first two requirements in the above definition imply that the image $f(B_r(p_n))$ is contained in the ball of radius $r + \varepsilon$ centered at p. This and

the third requirement imply (by Corollary 7.3.28) that the ball $B_r(p_n)$ in X_n lies within the Gromov–Hausdorff distance of order ε from a subset of X between the balls of radii $r - \varepsilon_n$ and $r + \varepsilon_n$ centered at p (here "between" means that the set contains one ball and is contained in the other).

If X is a length space, this remains true for the r-ball centered at p instead of the unknown subset; in other words, for every $r > 0$ the r-balls in X_n centered at p_n converge (with respect to the Gromov–Hausdorff distance) to the r-ball in X centered at p. In the general case (for nonlength spaces) such a simplification is not possible (see exercises below).

Exercise 8.1.3. Let $(X_n, p_n) \xrightarrow[GH]{} (X, p)$ and let X be a length space. Prove that $B_r(p_n) \xrightarrow[GH]{} B_r(p)$ for every $r > 0$.

Exercise 8.1.4. Show that the statement of the previous exercise fails without the assumption that X is a length space. To do that, construct a sequence $\{X_n\}$ of compact metric spaces converging to a compact metric space X such that no sequence of closed unit balls in X_n converges to a closed unit ball in X.

This property that the balls converge does not yet imply convergence of pointed spaces. The first requirement in Definition 8.1.1 includes additional information that puts the points p_n and p into a special position. Roughly speaking, not only should the balls converge, but also their distinguished central points should converge at the same time.

Exercise 8.1.5. Construct a compact metric space X and two points $p, q \in X$ such that for every $r > 0$ the balls $B_r(p)$ and $B_r(q)$ in X are isometric, but there is no isometry map from X to itself that maps p to q.

Hint: There are examples among finite spaces.

Given such X, p and q, let $X_n = X$ and $p_n = q$ for all $n \geq 1$. Prove that the sequence $\{(X_n, p_n)\}$ of pointed spaces does not converge to (X, p) despite the fact that $B_r(p_n) \xrightarrow[GH]{} B_r(p)$ for every $r > 0$.

Obviously if a sequence of pointed metric spaces converges to a pointed space (X, p), it also converges to its completion. We will only consider complete metric spaces as Gromov–Hausdorff limits. Then, similarly to the case of ordinary convergence, a Gromov–Hausdorff limit of pointed spaces is essentially unique.

Definition 8.1.6. We say that two pointed metric spaces (X, p) and (X', p') are *isometric* if there is an isometry $f : X \to X'$ such that $f(p) = p'$. Such a map f is called a *pointed isometry* from (X, p) to (X', p').

Theorem 8.1.7. *Let (X, p) and (X', p') be two (complete) Gromov-Hausdorff limits of a sequence $\{(X_n, p_n)\}_{n=1}^{\infty}$, and assume X is boundedly compact. Then (X, p) and (X', p') are isometric.*

Proof. We only outline the proof here; the details are similar to those of the proof of Theorem 7.3.30. Given an $r > 0$ and $\varepsilon > 0$, the map whose existence is guaranteed by Definition 8.1.1 can be converted into a correspondence $\mathfrak{R}_{r,\varepsilon}$ between sets of $Y_{r,\varepsilon} \subset X$ and $Y'_{r,\varepsilon} \subset X'$ such that: $Y_{r,\varepsilon}$ and $Y'_{r,\varepsilon}$ contain the balls of radius $r-\varepsilon$ and are contained in the balls of radius $r+\varepsilon$ centered at p and p', respectively; p and p' are in correspondence with each other; and $\operatorname{dis}\mathfrak{R}_{r,\varepsilon} < \varepsilon$.

Choosing one corresponding point for every point of Y_ε yields a map $f_{r,\varepsilon} : Y_{r,\varepsilon} \to Y'_{r,\varepsilon}$ that maps p to p' and has distortion $< \varepsilon$. Using the Cantor diagonal procedure first for $\varepsilon \to 0$ and then for $r \to \infty$, one can end up with a distance-preserving map from a dense subset of X to X', which extends to a distance-preserving map $f : X \to X'$ with $f(p) = p'$. Since f is distance-preserving, it maps every ball $B_r(p)$ in X to the corresponding ball $B_r(p')$ in X'.

In addition, the images of maps $f_{r,\varepsilon}$ are ε-nets in the respective subsets of X', and this implies that the balls $B_r(p')$ in X' are compact as well (compare with Exercise 7.3.31). Hence a similar distance-preserving map $f' : X' \to X$ exists. Due to compactness of balls, this implies that the restriction of f to $B_r(p)$ is an isometry onto $B_r(p')$ for every $r > 0$. Hence f is an isometry onto X'. □

In the sequel, we always assume that the pointed spaces under consideration are boundedly compact. As the next exercise shows, this property is inherited by limit spaces.

Exercise 8.1.8. Suppose that $(X_n, p_n) \xrightarrow[GH]{} (X, p)$ where the spaces X_n are boundedly compact and X is complete. Prove that X is boundedly compact.

Most properties of Gromov–Hausdorff convergence of compact spaces have their counterparts for pointed spaces. We collect the most important ones in the next two theorems.

Theorem 8.1.9. *Let $(X_n, p_n) \xrightarrow[GH]{} (X, p)$ where X_n are length spaces and X is complete. Then X is a length space.*

Proof. Repeat the proof of Theorem 7.5.1. □

The following is the version of compactness theorem for pointed convergence.

Theorem 8.1.10. *Let $\mathfrak{X}$ be a class of pointed metric spaces. Suppose that for every $r > 0$ and $\varepsilon > 0$ there exists an $N = N(r, \varepsilon)$ such that for every $(X, p) \in \mathfrak{X}$ the ball $B_r(p)$ in X admits an ε-net of no more than $N(r, \varepsilon)$ points. Then the class $\mathfrak{X}$ is precompact in the sense that any sequence of spaces in $\mathfrak{X}$ contains a converging subsequence.*

Proof. Similar to that of Theorem 7.4.15. Again, one has to apply the Cantor diagonal procedure twice, for $\varepsilon \to 0$ and for $r \to \infty$. □

8.2. Tangent and Asymptotic Cones

In this section we extend the notion of tangent and asymptotic cones discussed in Section 7.1, to abstract metric spaces.

Recall that for a metric space X and a $\lambda > 0$ one can consider a rescaled metric space λX which is the same set of points equipped with the original metric multiplied by λ. Rescaling a pointed space (X, p) naturally yields a pointed space $(\lambda X, p)$.

Definition 8.2.1. A pointed metric space (X, p) is called a *cone* if it is invariant under rescaling, i.e., if $(\lambda X, p)$ is isometric to (X, p) as a pointed space for any $\lambda > 0$.

Note that a cone is not necessarily a cone over a metric space as defined in Subsection 3.6.2.

8.2.1. Tangent cone. The tangent cone is a local notion which does not belong to large-scale geometry. Nevertheless we define it here because the definition is in some sense similar to that of the asymptotic cone.

Definition 8.2.2. Let X be a (boundedly compact) metric space, $p \in X$. A Gromov–Hausdorff limit of pointed spaces $(\lambda X, p)$ as $\lambda \to \infty$, if one exists, is called the *Gromov–Hausdorff tangent cone of* X *at* p.

As usual, the "limit as $\lambda \to \infty$" can be interpreted as the limit through any sequence of values for λ tending to infinity (which should exist and be the same for all sequences).

Note that the tangent cone is a pointed metric space. Its distinguished point (the natural ancestor of p) is called the origin or the apex of the cone. The tangent cone is indeed a cone in the sense that it is isometric to any dilatation of itself (via a pointed isometry). The tangent cone is a local invariant: it is determined by any small neighborhood of the point. More precisely, if U is a neighborhood of p in X, then the tangent cones of U and X at p are isometric. This follows immediately from the definition.

The tangent cone is indeed a cone in the sense of Definition 8.2.1. This follows from the following simple fact:

Exercise 8.2.3. Let $\{(X_n, p_n)\}_{n=1}^{\infty}$ be a sequence of pointed metric spaces, $(X_n, p_n) \xrightarrow[GH]{} (X, p)$. Prove that $(\lambda X_n, p_n) \xrightarrow[GH]{} (\lambda X, p)$ for any $\lambda > 0$.

Exercise 8.2.4. Let M be an n-dimensional Riemannian manifold. Prove that the tangent cone of M at any point exists and is isometric to $\mathbb{R}^n$.

Exercise 8.2.5. Prove that the tangent cone of a convex set, as defined in Subsection 7.1.1, is also the Gromov–Hausdorff tangent cone.

Remark 8.2.6. In Subsection 3.6.6 we introduced another local construction, the space of directions at a point. In fact, for "good" spaces the two constructions carry the same information; moreover the Gromov–Hausdorff tangent cone is nothing but the metric cone over the space of directions.

8.2.2. Asymptotic cone.

Definition 8.2.7. Let X be a (boundedly compact) metric space and $p \in X$. A Gromov–Hausdorff limit of pointed spaces $(\lambda X, p)$ as $\lambda \to 0$, if one exists, is called the *Gromov–Hausdorff asymptotic cone of* X, or a *cone of* X *at infinity.*

Proposition 8.2.8. *The asymptotic cone does not depend on the choice of the reference point* p.

Proof. Let (C, o) be the asymptotic cone of X defined with a point $p \in X$. Let $p' \in X$ be another point. We have to prove that $(\lambda X, p') \to (C, o)$ as $\lambda \to 0$. Let $f \colon \lambda X \to C$ be a map from Definition 8.1.1 of pointed convergence (for some $\varepsilon > 0$). Let us replace f by a map $f' \colon \lambda X \to C$ defined as follows: $f'(x) = f(x)$ for all $x \neq p'$, and $f'(p') = o$. We have moved the image of one point by a distance no greater than $\lambda|pp'| + \varepsilon$; hence the conditions from Definition 8.1.1 remain satisfied for p' instead of p with ε replaced by $2\varepsilon + \lambda|pp'|$ (check this!). Since ε and λ are arbitrarily small and $|pp'|$ is fixed, the desired convergence follows. □

Exercise 8.2.9. Prove that the asymptotic cone of a convex set, as defined in Subsection 7.1.2, is also its Gromov–Hausdorff asymptotic cone.

Exercise 8.2.10. Let X and Y be metric spaces and $d_{GH}(X, Y) < \infty$. Prove that, if X has an asymptotic cone, then Y has one too, and the two cones are isometric.

In particular, if a metric space X lies within a finite Gromov–Hausdorff distance from some cone Y, then Y is an asymptotic cone of X.

Exercise 8.2.11. Prove that the grid described in Subsection 7.1.3 has asymptotic cone isometric to $\mathbb{R}^2_1$.

The next exercise generalizes the previous one. There are further generalizations in Subsection 8.5.1.

Exercise 8.2.12. Consider a group $\mathbb{Z}^n$ and a symmetric finite set S of generators in it. Let d be the word metric associated with S, and $\|\cdot\|$ be

the norm on $\mathbb{R}^n$ whose unit ball is the convex hull of S. Prove that there exists a constant C (depending on S) such that

$$\|x - y\| \le d(x, y) \le \|x - y\| + C$$

for all $x, y \in \mathbb{Z}^n$.

In particular, the Gromov–Hausdorff distance between $(\mathbb{Z}^n, d)$ and the normed space $(\mathbb{R}^n, \|\cdot\|)$ is finite (not greater than C), and hence $(\mathbb{R}^n, \|\cdot\|)$ is the asymptotic cone of $(\mathbb{Z}^n, d)$.

The next exercise shows that the property of having a Gromov–Hausdorff asymptotic cone is not as common as one might think.

Exercise 8.2.13. Prove that the hyperbolic plane $\mathbb{H}^2$ does not have an asymptotic cone.

Hint: The sequence of rescaled balls $\frac{1}{n}B_n(p)$, $n \in \mathbb{N}$, $p \in \mathbb{H}^2$, is not uniformly totally bounded.

In plain words, the reason why the hyperbolic plane does not have an asymptotic cone is that its metric balls grow too fast when the radius goes to infinity. There are other constructions encoding asymptotic properties of such "fast growing" spaces and in other cases when Gromov–Hausdorff asymptotic cones are not applicable. One possible approach is to use a weaker type of limit than the Gromov–Hausdorff one. We do not discuss such generalized definitions in this book. Let us only mention that the "generalized" asymptotic cone of the hyperbolic plane is an infinite (and not locally finite) tree.

Another construction serving the same purposes is the ideal boundary of a space; cf. Subsection 5.3.3 for the case of hyperbolic plane. While the asymptotic cone is "the tangent cone at infinity" in some sense, the ideal boundary is a sort of "space of directions" at infinity.

8.3. Quasi-isometries

8.3.1. Definitions and first examples. Quasi-isometries are a large-scale analog of bi-Lipschitz maps. Two metric spaces are quasi-isometric if they are bi-Lipschitz equivalent up to a finite Gromov–Hausdorff distance. This is formally described as follows:

Definition 8.3.1. Metric spaces X and Y are said to be *quasi-isometric* if there exist metric spaces X' and Y' such that $d_{GH}(X, X') < \infty$, $d_{GH}(Y, Y') < \infty$, and the spaces X' and Y' are bi-Lipschitz homeomorphic (i.e., there exists a bi-Lipschitz homeomorphism between them).

We will soon see that X' and Y' can be chosen among subsets of X and Y, moreover, among *separated nets.*

Recall that a subset S of a metric space X is called a ρ-net if the Hausdorff distance between S and X is not greater than ρ.

Definition 8.3.2. Let X be a metric space. A set $S \subset X$ is called a *net* in X if the Hausdorff distance between X and S is finite. In other words, S is a ρ-net in X for a sufficiently large ρ.

A *separated net* is a net which is ε-separated for some $\varepsilon > 0$ (recall that being an ε-separated set means that $|xy| \geq \varepsilon$ for any two distinct points $x, y \in S$).

Every metric space X contains an ε-separated net for any given ε. Indeed, by Zorn's Lemma there exist an ε-separated set $S \subset X$ which is maximal by inclusion; i.e., if $S \subsetneq S'$, then S' is not ε-separated. Such S is an ε-net; indeed, if there is a point $x \in X$ with $\operatorname{dist}(x, S) \geq \varepsilon$, then the set $S' = S \cup \{x\}$ is again ε-separated, contrary to the maximality of S.

Let X, Y, X', Y' be as in Definition 8.3.1. Since $d_{GH}(X, X') < \infty$ and $d_{GH}(Y, Y') < \infty$, there are maps $f_1 \colon X \to X'$ and $f_2 \colon Y' \to Y$ with finite distortions and such that the images $f(X)$ and $g(Y')$ are nets in X' and Y, respectively. Let $g \colon X' \to Y'$ be a bi-Lipschitz homeomorphism and $\lambda = \max\{\operatorname{dil} g, \operatorname{dil} g^{-1}\}$. Define a map $f = f_2 \circ g \circ f_1$ from X to Y. Then

$$\frac{1}{\lambda}|xy| - C \leq |f(x)f(y)| \leq \lambda|xy| + C \tag{8.1}$$

for all $x, y \in X$ where $C = \operatorname{dis}(f_2) + \lambda \operatorname{dis}(f_1)$.

Definition 8.3.3. Let X and Y be metric spaces. A map $f \colon X \to Y$ is called a *quasi-isometry* if there are constants $C \geq 0$ and $\lambda \geq 1$ such that the inequality (8.1) holds for all $x, y \in X$.

Observe that the quasi-isometry $f \colon X \to Y$ constructed above possesses an additional property: its image $f(X)$ is a net in Y. Moreover the image of any net in X is a net in Y. To prove this, observe that each of the maps f_1, f_2 and g sends nets to nets.

Now choose a Δ-separated net $S \in X$ for a large enough Δ, namely, $\Delta > (2\lambda + 1)C$ where λ and C are from (8.1). Then (8.1) implies that

$$\frac{1}{2\lambda}|xy| \leq |f(x)f(y) \leq (\lambda + 1)|xy|$$

for all $x, y \in S$. Hence f is a bi-Lipschitz homeomorphism between S and $f(S)$. Note that S and $f(S)$ can be used instead of X' and Y' in Definition 8.3.1. Thus we have proved the following

Proposition 8.3.4. *For any metric spaces X and Y, the following three assertions are equivalent:*

(i) *X and Y are quasi-isometric;*

(ii) *there is a quasi-isometry $f\colon X \to Y$ whose image $f(X)$ is a net in Y;*

(iii) *X and Y contain bi-Lipschitz homeomorphic separated nets.* □

Corollary 8.3.5. *Being quasi-isometric is an equivalence relation.*

Proof. Symmetry is obvious from the definition. Transitiveness follows from (ii) in the above proposition. □

Example 8.3.6. Euclidean spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ are not quasi-isometric for $m \neq n$. To prove this, consider a separated net S in $\mathbb{R}^n$ and for every $R > 0$ denote by $N(R)$ the number of elements of S within the R-ball centered at the origin. Since S is a separated net, say, an ε-separated r-net, the $(\varepsilon/2)$-balls centered at points of S do not intersect one another, and the r-balls cover the space. Since the volume of a ball is proportional to the nth power of its radius, this leads to estimates for $N(R)$:

$$\frac{(R-r)^n}{r^n} \leq N(R) \leq \frac{(R+\varepsilon/2)^n}{(\varepsilon/2)^n},$$

or, in a simple form,

$$cR^n \leq N(R) \leq CR^n$$

for all large enough R, where c and C are some positive constants. This property of a net is obviously preserved by bi-Lipschitz maps, but no separated net in $\mathbb{R}^m$ satisfies it if $m \neq n$ (because for the same reasons the respective quantity for $\mathbb{R}^m$ lies between cR^m and CR^m).

Similarly, Euclidean spaces are not quasi-isometric to the hyperbolic plane (and hyperbolic spaces) because the areas (or volumes) of hyperbolic balls grows exponentially as radii go to infinity.

Exercise 8.3.7. Prove that hyperbolic spaces $\mathbb{H}^m$ and $\mathbb{H}^n$ are not quasi-isometric if $m \neq n$.

Hint: Supposing that they are quasi-isometric, prove that there is a *continuous* quasi-isometry from $\mathbb{H}^m$ to $\mathbb{H}^n$. Then consider this quasi-isometry restricted to a large sphere in $\mathbb{H}^m$ (assuming that $m > n$).

8.3.2. Groups and orbits. We already discussed metrics invariant under a group action in Section 3.3; below we revise this notion in a new context.

Let G be a group, X a set, and $\varphi : G \times X \to X$ an action of G on X (cf. Definition 3.3.5). As usual, we write $g(x)$ instead of $\varphi(g, x)$. A metric d on X is said to be G-invariant (under this action) if G acts by isometries,

i.e., if $d(g(x), g(y)) = d(x, y)$ for all $x, y \in X$, $g \in G$. We will consider only isometric actions (or, equivalently, only G-invariant metrics).

An action is said to be *co-compact* if the quotient space X/G is compact, and is said to be *co-bounded* if X/G is bounded. Obviously if X is boundedly compact (in particular, is a complete locally compact length space), then every co-bounded action of a group on X is co-compact. An *orbit* of an element $x \in X$ under an action of G, denoted by Gx, is by definition the set $\{gx : g \in G\}$.

Exercise 8.3.8. Prove that the orbit of a point is a net if and only if the action is co-bounded.

An important example of a group action is its action on itself by multiplication. Namely, one lets $X = G$ and defines the action by $g(h) = gh$ for all $g, h \in G$. Metrics on G invariant under this action are called *left-invariant* metrics (reflecting the fact that the group acts by left multiplication). If G is an abelian group, the prefix "left-" can be omitted. Note that word metrics discussed in Subsection 3.2.3 are left-invariant.

A left-invariant metric on G is determined by the distances from the identity of the group. Indeed, if we define $|g| = d(e, g)$, then $d(g, g_1) = |g^{-1}g_1|$. Conversely, a function $|\cdot|\colon G \to \mathbb{R}$ defines a metric if and only if

(1) $|e| = 0$, $|g| > 0$ for all $g \neq 0$;

(2) $|g^{-1}| = |g|$ for all g;

(3) $|g_1 g_2| \leq |g_1| \cdot |g_2|$.

(This is a trivial exercise.)

Note that these conditions on $|\cdot|$ are very similar but not identical to those in the definition of a norm on a vector space. For a word metric, the corresponding function $|\cdot|$ is the length of a shortest word representing a given element.

Here is a more general source of examples of left-invariant metrics on groups. Let X be an arbitrary set along with an action of G on it. Assume that the action is free, i.e. $g(x) \neq x$ for all $x \in X$ and $g \in G$ unless $g = e$. Then every G-invariant metric d_X on X and every point $x \in X$ determine a left-invariant metric d_G on G by the formula

$$d_G(g_1, g_2) = d_X(g_1(x), g_2(x)). \tag{8.2}$$

This can be interpreted as follows: one identifies G with the orbit Gx and uses the metric d restricted on Gx. The metric d_G defined by the above formula will be referred to as an *orbit metric*.

Exercise 8.3.9. Let X, d_X and G be as above, x_1 and x_2 be two points of X, d_1 and d_2 be the corresponding orbit metrics on G. Prove that the function $d_1 - d_2$ is bounded.

This construction becomes more interesting if we restrict ourselves to length metrics on X. Assume that the action of G is free and *totally discontinuous*, or, equivalently, the projection $p\colon X \to X/G$ is a covering map. (See Subsection 3.4.2 for a general discussion of coverings and length metrics.) Then there is a natural 1-1 correspondence between length metrics on X/G and G-invariant length metrics on X (cf. Proposition 3.4.16). Thus, given a length metric on X/G and a point $y \in X/G$, one can define a left-invariant metric on G by picking any point x such that $p(x) = y$ and taking the corresponding orbit metric. It is trivial that the result does not depend on the choice of x.

Since the orbit is just the set $p^{-1}(y)$, the group with this metric is isometric to $p^{-1}(y)$ with the metric restricted from X. If the action is co-compact, then $p^{-1}(y)$ is a net in X, so the Gromov–Hausdorff distance between X and the group is finite (so one can replace X by the group for all large-scale considerations).

Exercise 8.3.10. Prove the above statement: an orbit of a co-compact action is a separated net (of course, provided that the action is free, totally discontinuous and by isometries, and X is a length space).

Example 8.3.11. Word metrics (see Subsection 3.2.3) can be represented as orbit metrics of length spaces. To produce a word metric on a group G, let X be a Cayley graph (with the obviously defined action of G), and choose the identity of the group as the initial point of an orbit.

One can construct orbit metrics the other way round, starting from the quotient space. Let Y be a length space, and assume that Y is locally simply connected. Then there is a universal covering $p\colon X \to Y$. Furthermore, the fundamental group of Y naturally acts on X by isometries (as the group of deck transformations). Thus every length metric on Y and a point $y \in Y$ determine a left-invariant metric on the fundamental group $\pi_1(Y, y)$, namely, the metric of X restricted to the orbit $p^{-1}(y)$. Here is another description of this metric:

Definition 8.3.12. Let Y be as above, $y \in Y$, $g \in \pi_1(Y, y)$. The *length* of g is defined by

$$\operatorname{length}(g) = \inf\{L(\gamma) : \gamma \text{ is a loop representing } g \text{ in } \pi_1(Y, y)\}.$$

The distance between two elements $g_1, g_2 \in \pi_1(Y, y)$ is the length of $g_1^{-1}g_2$.

Exercise 8.3.13. Prove that the above definition defines the same metric on $\pi_1(Y, y)$ as in the construction with the universal covering.

Remark 8.3.14. One can apply the same construction to other coverings, not only to the universal one; however the covering must be regular. (This is necessary in order to represent Y as the quotient space of the group of deck transformations.) In this case, the group of deck transformation is a factor-group of $\pi_1(Y)$.

We will soon see that all possible orbit metrics of co-compact actions on a given (finitely generated) group are bi-Lipschitz equivalent to one another. In other words, the quasi-isometry class of a metric on a group depends only on the group, not on a length space and an action used to define the metric. This yields the following remarkable corollary: *if two compact length spaces have isomorphic fundamental groups, then their universal coverings are quasi-isometric.* Indeed, each universal covering has finite Gromov–Hausdorff distance from the fundamental group equipped with an orbit metric, and two orbit metrics are bi-Lipschitz equivalent (as we will prove).

Since all metrics on a fixed group are bi-Lipschitz equivalent, we can talk about quasi-isometries between abstract finitely generated groups without referring to a particular metric. Namely, two groups are quasi-isometric if they are quasi-isometric when equipped with arbitrary word metrics.

Example 8.3.15. The free groups F_2 and F_3 with two and three generators (equipped, say, with word metrics) are quasi-isometric. (Compare this with the fact that $\mathbb{Z}^m$ and $\mathbb{Z}^n$ are not quasi-isometric if $m \neq n$, as we have seen in Example 8.3.6.)

To prove this, recall that F_2 is the fundamental group of a bouquet of two circles (say, of unit length). This bouquet X admits a two-sheeted covering by a graph Y consisting of three circles S_1, S_2, S_3 connected as a chain: S_1 has one common point with S_2, S_2 has one common point with S_3, and $S_1 \cap S_2 = \emptyset$. Metrically, Y is a circle of length 2 glued with two circles of unit length; smaller circles are attached to opposite points of the larger one. Constructing the covering map from Y to X is left as an exercise.

Since Y is a covering space for X, the universal covering of Y coincides with the universal covering of X. This covering space is quasi-isometric to both fundamental groups $\pi_1(X) \cong F_2$ and $\pi_1(Y) \cong F_3$. (The latter group is F_3 because Y is homotopy equivalent to a bouquet of three circles. To see this, just contract of the edges.) Thus F_2 and F_3 are quasi-isometric.

Exercise 8.3.16. Construct an explicit quasi-isometry from F_2 onto a net in F_3.

Exercise 8.3.17. Prove that all free groups F_n, $n \in \mathbb{N}$, are quasi-isometric.

Now we turn to the proof of the above claimed equivalence of orbit metrics. We begin with the case of word metrics.

Proposition 8.3.18. *Let G be a finitely generated group. Then all word metrics on G are bi-Lipschitz equivalent to one another.*

Proof. Let $|\cdot|_1$ and $|\cdot|_2$ be the distance functions of the identity in two word metrics defined by generating sets S_1 and S_2, respectively. Recall that the sets S_1 and S_2 must be finite. Let $g_1 \dots g_n$ be a shortest word in generators from S_1 representing a given element $g \in G$. Then $|g|_1 = n$ and

$$|g|_2 = |g_1 \dots g_n|_2 \le |g_1|_2 \cdot \dots \cdot |g_n|_2 \le C_1 n = C_1 |g|_1,$$

where $C_1 = \max_{h \in S_1} |h|_2$. (Note that we have not yet used that $|\cdot|_2$ is a word metric.) Similarly, $|g|_1 \le C_2 |g|_2$ for some constant C_2 not depending on g. Thus $|\cdot|_1$ and $|\cdot|_2$ are bi-Lipschitz equivalent. □

Theorem 8.3.19. *Let G be a finitely generated group and d be an orbit metric of a free co-compact action of G by isometries on a length space X. Then d is bi-Lipschitz equivalent to a word metric.*

In particular, all such orbit metrics on G are bi-Lipschitz equivalent to one another.

Corollary 8.3.20. *All length spaces X admitting a free totally discontinuous co-compact action of a given group G are quasi-isometric to one another, and are quasi-isometric to the group G equipped with any word metric.*

In particular, if Y is a length space and X its universal covering with the metric lifted from Y, then X is quasi-isometric to the group $\pi_1(Y)$ equipped with any word metric.

Proof of the theorem. Let $|\cdot|$ denote the distance from the identity in the metric d. If $|\cdot|_w$ is a word metric, then $|\cdot| \le C|\cdot|_w$ for some constant C, similarly to the previous proof. Thus it suffices to prove that $|\cdot|_w \le C|\cdot|$ for some constant C and *some* word metric $|\cdot|_w$.

Let $x \in G$ be the point whose orbit is used to define the metric d (i.e., $d(g,h) = d_X(gx, hx)$ for all $g, h \in G$). Since the action is co-compact, X is locally compact and complete and the orbit Gx is a separated net. Therefore every metric ball in X contains only finitely many elements of the orbit. Hence every ball in (G,d) is a finite set.

Let D be a number such that the orbit Gx is a D-net in X, and let r be so large that $r > 2D + 1$ and the r-ball in (G,d) centered at the identity contains a set of generators. We may assume that this ball itself is chosen as the set of generators S defining the word metric $|\cdot|_w$. (We have an option to choose a word metric to be compared with d; then the result will follow for any word metric because all word metrics are already proven to be equivalent.)

Pick a $g \in G$ and let γ be a shortest path in X connecting x to gx. We are going to show that $|g|_w \leq C|g| = C \cdot L(\gamma)$ for a constant C not depending on g. Divide γ into intervals of length 1 or less by points $x_1, \dots, x_n$, where $n \leq L(\gamma) \leq n+1$. For every $i = 1, \dots, n$, find a $y_i \in Gx$ such that $d_X(x_i, y_i) \leq D$. In addition, set $y_0 = x$ and $y_{n+1} = Gx$. Let $g_i \in G$ be such that $g_i x = y_i$ (note that $g_0 = e$ and $g_{n+1} = g$), and set $h_i = g_i g_{i-1}^{-1}$. Then $h_i x_{i-1} = x_i$; hence $|h_i| = d_X(x_{i-1}, x_i) \leq 2D+1$. Therefore h_i belongs to our set of generators (recall that it is just the r-ball in G, and $r > 2D+1$). On the other hand, the product (word) $h_1 \dots h_{n+1}$ equals g; hence $|g|_w \leq n+1$.

Thus $|g|_w \leq L(\gamma) + 1 = |g| + 1$. Since the distances $|g|$ are separated away from zero (i.e., there is an $\varepsilon > 0$ such that $|g| \geq \varepsilon$ for all $\varepsilon > 0$), this estimate can be written in the desired form $|g|_w \leq C|g|$, namely, $|g|_w \leq (1 + 1/\varepsilon)|g|$. □

Remark 8.3.21. One can remove the condition that the action is free from the theorem. The only problem is that an orbit metric is no longer a metric; it is only a semi-metric. As a consequence, it is not bi-Lipschitz equivalent to word metrics, but it is still quasi-isometric to them. The same proof works with obvious modifications.

The condition that a given metric is an orbit metric of some action seems hard to verify and somewhat restrictive (it is not a "large-scale" one, after all). There are various possible replacements for it; one of such conditions is given in the following exercise.

Exercise 8.3.22. Prove the following statement which is slightly more general than the above theorem. Let d be a left-invariant metric on a finitely generated group G such that every ball of d is a finite set, and there is a constant $C > 0$ satisfying the following: for any $x, y \in G$ there is a $z \in G$ with $d(x,z) < \frac{1}{2}d(x,y) + C$ and $d(y,z) < \frac{1}{2}d(x,y) + C$. Then d is bi-Lipschitz equivalent to the word metric.

8.4. Gromov Hyperbolic Spaces

In this section we will deal with coarse geometry; this means that we will study our metric spaces on the large scale of distances (large-scale geometry), neglecting all phenomena related to distances smaller than a certain value. From this viewpoint, a space can be substituted by a discrete subset forming an ϵ-net and carrying discrete topology; hence all local topology is irrelevant for our considerations. A model example of this situation is the universal cover of a compact space X; the universal cover can be substituted by the fundamental group of X with the induced metric. We will be mainly discussing *δ-hyperbolicity*—a version of "coarsely negative curvature" introduced by Gromov. Loosely speaking, δ-hyperbolicity reflects

large-scale features of hyperbolic planes (or, more generally, of spaces of curvature bounded above by a negative number; such spaces are discussed in Chapter 9). (One notices that these properties have to do with large distances only.)

8.4.1. δ-hyperbolicity. Let us recall that a triangle $\triangle a_1a_2a_3$ in the hyperbolic plane has the following property: there exists a "center", that is, a point c such that all distances from this point to the sides of the triangle are less than 1 (see Subsection 5.3.5). Thus from our "coarse viewpoint" the triangle is "slim": it looks as if all its sides pass through one point c, and the whole triangle looks like a bouquet of three segments $[ca_1]$, $[ca_2]$, $[ca_3]$. In other words, every side of a triangle belongs to 1-neighborhood of the union of the two other sides. One can use this property to define large-scale hyperbolicity:

Definition 8.4.1. A length space (X, d) with a strictly intrinsic metric d is said to be δ-hyperbolic, $\delta \geq 0$, if all triangles in (X, d) satisfy the following property: each side of a triangle belongs to the δ-neighborhood of the union of the two other sides.

Throughout this section by default we assume that *all constructions take place in a δ-hyperbolic space (X, d)*.

Clearly if a length space is δ-hyperbolic for a particular δ, it is also δ'-hyperbolic for all $\delta' \geq \delta$. Following our large-scale ideology, we will not be concerned with a particular value of δ; a space is said to be *Gromov hyperbolic* if it is δ-hyperbolic for some $\delta > 0$.

Note that, in particular, this definition implies that two shortest paths between the same points in a δ-hyperbolic space are within distance δ from each other (to see this, one just regards the bi-gone formed by the shortest path as a degenerate triangle with two coinciding vertices). Moreover, the following proposition holds (we leave the proof to the reader as a trivial exercise):

Lemma 8.4.2. *A shortest path $[ab]$ belongs to the $(\delta + d(b,c))$-neighborhood of a shortest path $[ac]$.*

For three points p, q, r, we will use the notation

$$(p,q)_r = \frac{1}{2}(d(r,p) + d(r,q) - d(p,q)).$$

Thus $(p,q)_r$ measures "to what extent the triangle inequality for the triangle $\triangle pqr$ is far from being an equality". Let us say that a triangle $\triangle pqr$ is δ-almost degenerate if the triangle inequality is almost an equality: $(p,q)_r \leq \delta$.

Consider a point p in the side $[bc]$ of a triangle $\triangle abc$ in a Gromov hyperbolic space. It immediately follows from the definition that at least

one of the triangles $\triangle apb$ and $\triangle apc$ is δ-almost degenerate; in other words, the following inequality holds:

$$0 \geq \min((a,b)_p, (a,c)_p) - \delta. \tag{8.3}$$

Now we can see that in a δ-hyperbolic space $(b,c)_a$ is approximately the distance from a to $[bc]$. Indeed, by the triangle inequality we obviously have $(b,c)_a \leq d(a,[bc])$. Moving p continuously from b to c along $[bc]$, we find a value for p such that both triangles $\triangle pac$ and $\triangle pbc$ are δ-almost degenerate. Hence

$$2\delta \leq d(b,p) + d(p,a) - d(a,b),$$
$$2\delta \leq d(c,p) + d(p,a) - d(a,c).$$

Adding these inequalities (and using

$$d(b,p) + d(p,c) = d(b,c) = d(a,c) + d(a,b) - 2(b,c)_a,$$
$$d(a,[bc]) \leq d(a,p)),$$

we get $d(a,[bc]) \leq (b,c)_a + 2\delta$, and hence

$$(b,c)_a \leq d(a,[bc]) \leq (b,c)_a + 2\delta. \tag{8.4}$$

Reasoning analogously to the argument used to obtain inequalities (8.4), one proves the following useful lemma:

Lemma 8.4.3. *Let $[ab]$ and $[ac]$ be two shortest paths. Let $b' \in [ab]$ and $c' \in [ac]$ be such that $d(a,b') = d(a,c') \leq (b,c)_a$. Then $d(b',c') \leq 3\delta$. Moreover, there is a point $p \in [bc]$ such that $[bp]$ lies in δ-neighborhood of $[ba]$, and $[cp]$ lies in δ-neighborhood of $[ca]$.*

Exercise 8.4.4. Prove the lemma.

Though easy to prove, geometrically this lemma is quite striking. Consider two shortest paths $[ab]$ and $[ac]$ of length R starting from the same point. Let us think of this point a as a reference point (center), and R a large number (compared to δ). For instance, let us assume now that the distance between b and c is also R. In the Euclidean plane such shortest paths would form an angle of 60 degrees. However, in a δ-hyperbolic space they go very close (within distance 3δ) to each other until at least half-way! Moreover, a shortest path $[bc]$ connects b and c in a pretty weird way: first it goes (within distance δ) with the radial segment $[ba]$; then at some point $[ba]$ and $[ca]$ get very close to each other, and the shortest path $[bc]$ "turns back" and goes within distance δ along with $[ac]$. We suggest that the reader make a sketch of this situation, for we will exploit it a lot (in particular in the proof of the Morse Lemma).

Using (8.4), which has been derived solely from the inequality (8.3), it is easy to show that (8.3) alone implies Gromov hyperbolicity:

Exercise 8.4.5. Show that if the inequality (8.3) is satisfied for all triangles $\triangle a_1a_2a_3$ and all choices of $p \in [a_2a_3]$ for a strictly intrinsic metric d on X, then (X, d) is Gromov hyperbolic.

Consider a triangle $\triangle abc$ and a point p (in a δ-hyperbolic space). Since $[bc]$ is contained in the δ-neighborhood of the union of $[ab]$ and $[ac]$, we have $d(p, [bc]) \geq \min(d(p, [ab]), d(p, [ac])) - \delta$. Combining this with (8.4), we obtain that

$$(b, c)_p \geq \min((a, b)_p, (a, c)_p) - 3\delta. \tag{8.5}$$

Notice that the left-hand side of the inequality (8.3) (i.e., zero) can be represented as $(b, c)_p$ (for in (8.3) p is supposed to belong to $[bc]$), and hence (8.5) in its turn implies (8.3) with $\delta' = 3\delta$ in place of δ.

Now we have arrived at Gromov's original definition of Gromov hyperbolic spaces:

Definition 8.4.6. A metric space (X, d) is said to be Gromov hyperbolic if there exists a $\delta \geq 0$ such that for every four points $a, b, c, p \in X$, the following inequality holds:

$$(b, c)_p \geq \min((a, b)_p, (a, c)_p) - \delta.$$

This elegant definition has an important advantage: it does not involve shortest paths (sides). It uses only distance measurements and defines Gromov hyperbolic *metric* spaces (which can even fail to be length spaces). However, it is more difficult to visualize, and this is the reason why we began with an alternative (though less general) definition. Definition 8.4.6 is also sometimes harder to verify (because (8.3) is a particular case of (8.5)).

The reader can see that solving the following exercises, which state that spaces with strictly negative curvature are Gromov hyperbolic, whereas Euclidean spaces (except the real line) are not:

Exercise 8.4.7. Give a $\delta = \delta(k)$ such that the hyperbolic plane of curvature k is δ-hyperbolic.

Exercise 8.4.8. Show that a complete simply connected length space of curvature bounded above by $-k$ in the large, where $k > 0$, is a δ-hyperbolic space for all $\delta \geq 2/\sqrt{k}$. Can you suggest a better value for δ?

Exercise 8.4.9. Prove that the Euclidean plane is not Gromov hyperbolic.

To get more insight into the geometric meaning of Definition 8.4.6, let us look what it means for a 4-point space:

Lemma 8.4.10. *Let X be a 4-point 0-hyperbolic space. Then X is isometric to the set of leaves of a tree (with at most six vertices). Here the tree carries*

an intrinsic metric (it is just a finite graph, by the way isometric to an H-shaped subset of Euclidean plane formed by five segments, with its intrinsic metric).

Lemma 8.4.11. *Let X be a 4-point δ-hyperbolic space. Then X can be mapped to the set of leaves of a tree (with at most six vertices) so that all (the six) distances are distorted by no more than 2δ.*

We leave the proofs to the reader as an exercise. To better visualize the statement of Lemma 8.4.11, the following consideration maybe helpful. Let a, b, c, d be four points (in a δ-hyperbolic length space). Choose a point e in the intersection of δ-neighborhoods of $[ab]$, $[ac]$, and $[bc]$; similarly choose f in the intersection of δ-neighborhoods of $[ad]$, $[ac]$, and $[dc]$. Let $d(a,e) \leq d(a,f)$. Then all shortest paths connecting a, b, c, d lie in 2δ-neighborhood of the tree formed by shortest segments $[ae]$, $[ef]$, $[fc]$, $[be]$, $[cf]$.

Exercise 8.4.12. Show that a 0-hyperbolic length space is a topological tree (every two points are ends of exactly one subset homeomorphic to a segment).

Furthermore, every Gromov hyperbolic space "on the large scale looks like a tree". The idea becomes clear from the following trivial observation: if (X, d) is δ-hyperbolic, then $(X, c \cdot d)$, $c > 0$, is (δ/c)-hyperbolic (check it!). To formalize this we need the following definition:

Definition 8.4.13. Let $K = \{p_1, p_2, \ldots, p_n\}$ be a finite metric space. K is said to be a *finite subcone* at infinity for X if there is a sequence $N_i \to \infty$, and a sequence of n-tuples of points $q_1^i, q_2^i, \ldots, q_n^i \in X$ with

$$N_i \frac{d(p_l^i, p_m^i)}{d(q_l^i, q_m^i)} \to 1 \quad \text{for all} \quad l, m \in \{1, 2, \ldots, n\} \quad \text{as} \quad i \to \infty.$$

A metric space Y is said to be a *subcone* at infinity for X if every finite subset of Y is a finite subcone at infinity for X.

Lemma 8.4.14. *Let K be a finite subcone at infinity for X. Then K is isometric to a subset of a finite tree with intrinsic metric on it. If Y is a subcone for X at infinity, then Y is a topological tree.*

The proof is an obvious corollary of the following important proposition, which we leave as an exercise. This proposition characterizes Gromov hyperbolic spaces by generalizing Lemma 8.4.11 from 4-point sets to all finite subsets of a Gromov hyperbolic space:

Lemma 8.4.15. *Let X be a Gromov hyperbolic space, and $n \in \mathbb{N}$. There exists a $C \geq 0$ such that any n-element subset of X can be mapped to the set of leaves of a finite tree so that all distances are distorted by no more than C.*

8.4.2. Bi-Lipschitz equivalence, quasi-geodesics, and Morse lemma. It follows immediately from Definition 8.4.6 that if a metric space is within finite Gromov-Hausdorff distance from a Gromov hyperbolic space, it is also Gromov hyperbolic (perhaps for a different value of δ). Indeed, bounded additive change in distance can be compensated by appropriately increasing δ in the inequality (8.5) (we suggest that the reader carry out a rigorous argument). Note that if one looks at Definition 8.4.1, it is not at all obvious that the property of being Gromov hyperbolic persists if the distances are changed by a uniformly bounded additive function.

Amazingly enough, for length spaces (and even for a more general class of *quasi-geodesic spaces*) Gromov hyperbolicity also persists under bi-Lipschitz equivalence. Recall that two metric spaces (X, d) and (X_1, d_1) are said to be bi-Lipschitz equivalent if there exist a bijective map $f : X \to X_1$ and a positive constant C such that

$$C^{-1}d(x, y) \leq d_1(f(x), f(y)) \leq Cd(x, y)$$

for all $x, y \in X$.

Theorem 8.4.16. *Let (X_1, d_1) be a length space bi-Lipschitz equivalent to a Gromov hyperbolic length space (X_1, d_1). Then (X_1, d_1) is also Gromov hyperbolic.*

Corollary 8.4.17. *Gromov hyperbolicity persists under quasi-isometries.*

Roughly speaking, this theorem asserts that if we multiply an intrinsic metric of a Gromov hyperbolic length space by a bounded factor and obtain a new length space, it is also Gromov hyperbolic. Thus we have a huge source of examples of Gromov hyperbolic spaces: we can begin with the hyperbolic plane and multiply its metric by a function bounded between two positive constants (note that this new metric can now have a lot of positive curvature!) We suggest that the reader have another look at Definition 8.4.6 to realize that this is a highly nontrivial phenomenon: the definition requires that certain inequalities for distance functions hold up to an additive constant δ; at the same time, we are allowed to *multiply* distances, even though the latter can be arbitrarily large, by a bounded factor.

The proof of Theorem 8.4.16 is based on a very important phenomenon: shortest paths in a Gromov hyperbolic space are stable (up to a uniformly bounded displacement) under bi-Lipschitz changes of metric. This property turns out to be equivalent to Gromov hyperbolicity. We precede it with the following definition.

Definition 8.4.18. A path γ (in a length space (X, d)) is called C-quasi-geodesic if $L(\gamma)_{[s,t]} \leq C \cdot d(\gamma(s), \gamma(t))$ for all s, t in the domain of γ. In other words, the length of every segment of γ is at most C times longer than the distance between its endpoints.

Of course, 1-quasi-geodesics are just shortest paths.

Definition 8.4.19. A length space is said to be quasi-geodesically stable if, for every c, there exists M with the following property: If $\gamma : [a, b] \to M$ is a C-quasi-geodesic segment and $[\gamma(a)\gamma(b)]$ a shortest path between its endpoints, then (the image of) γ belongs to the M-neighborhood of $[\gamma(a)\gamma(b)]$.

Now Theorem 8.4.16 is immediately implied by the following assertion, known as the Morse Lemma:

Theorem 8.4.20 (Morse lemma). *Gromov hyperbolic spaces are quasi-geodesically stable. More precisely, let (X, d) be a δ-hyperbolic length space and $C \in \mathbb{R}$. Then there exists a constant $M = M(\delta, C)$ with the following property: If $\gamma : [a, b] \to M$ is a C-quasi-geodesic segment and $[\gamma(a)\gamma(b)]$ a shortest path between its endpoints, then (the image of) γ belongs to the M-neighborhood of $[\gamma(a)\gamma(b)]$.*

In Euclidean plane every semi-circle is $(\pi/2)$-quasi-geodesic; and a semi-circle of a radius R gets as far as R away from the segment between its endpoints. In contrast, for a Gromov hyperbolic world the Morse Lemma asserts that if we want to travel in such a way that our path is never too long a detour (not longer than c times the distance between points), we are confined to travel within a bounded distance M from some shortest path! Another striking corollary of the Morse Lemma is the following fact: one cannot change large-scale behavior of geodesics in the hyperbolic plane by bounded distortions of metric: if a metric is multiplied by a function bounded away from 0 and infinity, new lines stay within a bounded distance from hyperbolic lines.

As a matter of fact, the property expressed by Theorem 8.4.20 is equivalent to Gromov hyperbolicity (see [**BM**]) as follows:

Definition 8.4.21. A length space (X, d) is said to be *Gromov hyperbolic* if it is quasi-geodesically stable.

Proof of the Morse Lemma. The proof consists of several steps, and we advise the reader to try to use them as exercises before reading our proofs. Though the proof contains ugly estimates, they are needed only to formally show that certain distances play no role for our large-scale consideration. For each distance, it is very useful to realize whether it is "of order δ" or it can become comparable to the size of the large object we are looking at.

First we want to look at the following situation: there are two points b and c on the boundary of a ball of radius R (centered at a point a), and the points are connected by a path γ never entering the ball. As usual, we think

of R as a large number (when compared to δ). We are going to show that if the distance $d(b,c)$ is not too small, then γ is much longer than the distance $d(b,c)$ (and hence cannot be a C-quasi-geodesic for a small C). The reason is that large spheres remain "rather convex" (similarly to the hyperbolic plane, and unlike the Euclidean case when large spheres get more and more flat). More precisely:

Lemma 8.4.22. *In the above notation, let L be the length of γ. Assume that $L \geq R \geq k^2\delta$ for a natural k. Then $d(a,b) \leq 10k^{-1}L$, and in particular γ cannot be a $10k^{-1}$-quasi-geodesic.*

Proof. Choose points $b_0 = b = \gamma(t_0), b_1 = \gamma(t_1), \ldots, b_n = c = \gamma(t_n)$ along γ so that the length of each segment $\gamma|_{[t_i,t_i+1]}$ is between $\frac{R}{k}$ and $\frac{R}{2k}$. Obviously the number of points n is bounded by

$$n \leq \frac{2kL}{R}. \tag{8.6}$$

Let $b'_i \in [b_i a]$ with $d(b'_i, a) = R(1-\frac{2}{k})$. By the triangle inequality $d(b'_i, a) \leq (b_i, b_{i+1})_a$. By Proposition 8.4.3, segments $d(a'_i, a'_{i+1}) \leq 3\delta$. Now consider the following broken line connecting b and c:

$$b = b_0, b'_0, b'_1, b'_2, \ldots, b'_n, b_n = c.$$

The length of segments $[bb'_0]$ and $[b'_n b]$ is $\frac{2R}{k}$, and the length of each segment $[b'_i b'_{i+1}]$ is at most 3δ. Thus the total length of the broken line, which gives us an upper bound on the distance from b to c, is at most $\frac{4R}{k} + 3n\delta$. Combining this with (8.6), we get

$$d(b,c) \leq \frac{4R}{k} + \frac{6kL}{R}\delta.$$

Using that $L \geq R$ and $R \geq k^2\delta$ (and hence $\frac{\delta}{R} \leq \frac{1}{k}$), we re-write this as

$$d(b,c) \leq \frac{4L}{k} + 6kL\frac{1}{k} = \frac{10}{k}L.$$

□

The next proposition is motivated by the following consideration. Let γ_1, γ_2 be two hyperbolic rays orthogonal to the segment $[\gamma_1(0), \gamma_2(0)]$. Assume that the length of this segment is, say, 10: $d(\gamma_1(0), \gamma_2(0)) = 10$. Then the distance $d(\gamma_1(T_1), \gamma_2(T_2))$ is at least $T_1 + T_2$.

Exercise 8.4.23. Prove this assertion.

Lemma 8.4.24. *Let $b, c, b_1, c_1 \in X$ be such that $d(b,b_1) = R_1$, $d(c,c_1) = R_2$, $d(b_1,c_1) \geq 4\delta$, and b_1, c_1 are closest points to b, c in the segment $[b_1,c_1]$. Then $d(b,c) \geq R_1 + R_2 - 4\delta$.*

Proof. Let us make a preliminary observation: the intersection of $[b_1c_1]$ and 2δ-neighborhoods of $[b_1b]$ is contained in a ball of radius 2δ centered at b_1 (this immediately follows from the assumption that b_1 is a point closest to b in $[b_1c_1]$). Similarly, the intersection of $[c_1b_1]$ and 2δ-neighborhoods of $[c_1c]$ is contained in a ball of radius 2δ centered at c_1 .

By definition of δ-hyperbolicity, $[b_1c]$ is contained in the δ-neighborhood of $[bb_1] \cup [bc]$. Hence the δ-neighborhood of $[b_1c]$ is contained in the 2δ-neighborhood of $[bb_1] \cup [bc]$. Now $[b_1c_1]$ is contained in the δ-neighborhood of $[b_1c] \cup [c_1c]$. Hence $[b_1c_1]$ is contained in the 2δ-neighborhood of $[b_1b] \cup [bc] \cup [cc_1]$. Since the intersection of $[b_1c_1]$ with $[b_1b] \cup [cc_1]$ is contained in the union of two balls of radius 2δ, whereas $d(b_1c_1) \geq 4\delta$, we see that $[b_1c_1]$ intersects the 2δ-neighborhood of $[bc]$. Let $p \in [bc]$ be such that $d(p,q) \leq 2\delta$ for some $q \in [b_1c_1]$. Then

$$d(b,c) = d(b,p) + d(p,c) \geq d(b,q) - 2\delta + d(c,q) - 2\delta \geq d(b,b_1) + d(c,c_1) - 4\delta,$$

where the last inequality uses $d(b,b_1) \leq d(b,q)$ and $d(c,c_1) \leq d(c,q)$, since b_1 and c_1 are closest points to b and c respectively in $[b_1c_1]$. □

Now we are ready to prove the Morse Lemma (Theorem 8.4.20). Consider a C-quasi-geodesic γ parameterized by arc length and connecting $p = \gamma(0)$ and $q = \gamma(T)$. Let $2R = \max_{t\in[0,T]} d(\gamma(t),[pq])$, and choose τ with $d(\gamma(\tau),[pq]) = 2R$.

Reasoning by contradiction, assume that $R > k^2\delta$, where $k > 20C + 8$ is a natural number.

Let $[t',t'']$ be the smallest interval containing τ and such that $d(\gamma|[t',t''],[pq]) \geq R$ (in particular, $d(\gamma(t'),[pq]) = d(\gamma(t''),[pq]) = R$). It is obvious that $|t''-t'|$ (which is equal to $L(\gamma,t',t'')$) is at least $2R$. Choose $t_0 = t' > t_1 > t_2 \cdots > t_n = t''$ so that $R/2 \leq t_i - t_{i-1} \leq R$. Then $|t''-t'| \geq nR/2$. Denote $b = b^0 = \gamma(t') = ga(t_0)$, $b^i = \gamma(t_i)$. Let b_1^i be a point closest to b^i in $[pq]$.

By Lemma 8.4.24 one has $d(b_1^i, b_1^{i+1}) \leq 4\delta$(recall that $R > 8\delta$). Thus $d(\gamma(t'),\gamma(t'')) \leq 4R + 4n\delta$, and therefore

$$C \geq \frac{|t''-t'|}{d(\gamma(t'),\gamma(t''))} \geq \frac{nR}{8R+8n\delta} \geq \frac{n}{9}.$$

Hence $n \leq 9C$.

Now set $a = b_1^0$; then $d(a,\gamma(t')) = R$. Obviously $\gamma|[t't'']$ lies outside a ball of radius R centered at a. Let c be the point in a segment $[a\gamma(t'')]$ at distance R from a. Since $d(b_1^n,\gamma(t'') = R)$ and $d(b_1^0,b_1^n) \leq 4n\delta \leq 36C\delta$, by the triangle inequality

$$d(c,\gamma(t'')) \leq 36C\delta < d(\gamma(t'),\gamma(t'')) \tag{8.7}$$

and

$$(8.8) \qquad d(b,c) \geq d(\gamma(t'),\gamma(t'')) - 36C\delta > \frac{1}{2}\, d(\gamma(t'),\gamma(t'')).$$

(Note that these inequalities are very crude; but all we want is to show that $[c\gamma(t'')]$ is short and plays no role in our large-scale considerations.)

Then a path consisting of $\gamma|_{[t't'']}$ continued by a segment $[\gamma(t'')c]$ also lies outside the ball of radius R centered at a. Applying Lemma 8.4.22 to this path, we can estimate the length of $\gamma|_{[t't'']}$ as

$$|t''-t'| \geq \frac{k}{10}\, d(b,c) - d(\gamma(t''),c).$$

Combining this with (8.7) and (8.8), we get

$$|t''-t'| \geq (\frac{k}{20} - 1)\, d(\gamma(t'),\gamma(t'')).$$

Since $k/20 - 1 > C$, this contradicts the assumption that γ is a C-quasi-geodesic. □

8.4.3. Hyperbolic groups. A *hyperbolic group* is finitely-generated group which is a Gromov hyperbolic metric space with respect to some word metric. Then such a group happens to be Gromov hyperbolic for *any* word metric, and hence hyperbolicity is an algebraic property of a group. One can very well think of the notion of a hyperbolic group as a generalization of basic features of fundamental groups of negatively curved compact manifolds. Our goal is not to give a systematic introduction to the theory of hyperbolic groups; we will only discuss some of their main properties. In particular, we will give three different definitions of hyperbolic groups and verify them for the fundamental groups of negatively curved manifolds.

Let Γ be a finitely generated group; that is, there is a finite symmetrical subset $S \subset \Gamma$ such that every element of Γ can be represented as a finite product of elements of S (here symmetry means that $s \in S$ implies $s^{-1} \in S$). Recall that the word distance between two elements $\gamma_1, \gamma_2 \in \Gamma$ is defined as the smallest k such that $\gamma_1\gamma_2^{-1}$ can be represented as $\gamma_1\gamma_2^{-1} = s_1 s_2 \dots s_k$, $s_i \in S$. Of course, this is not an intrinsic metric, for it takes only integer values, and topologically Γ with this metric is just a discrete space. However, Γ with a word metric is isometrically embedded into a Cayley graph $C(\Gamma)$ of Γ. Recall that the latter has Γ as its set of vertices, and two vertices γ_1 and γ_2 are connected by an edge (of length 1) if and only if $\gamma_1 s = \gamma_2$ for some $s \in S$. Then the word distance between two elements of Γ is the natural intrinsic distance between them in the Cayley graph. Of course, the Gromov-Hausdorff distance between Γ and $C(\Gamma)$ is finite, and thus they

share all large-scale properties. We will usually prefer to deal with $C(\Gamma)$ since it is a length space.

Definition 8.4.25. A group Γ is said to be *hyperbolic* (with respect to some finite generating set S) if the Cayley graph $C(\Gamma)$ (equivalently, just Γ) is a Gromov hyperbolic metric space.

Recall that any two word metrics are bi-Lipschitz equivalent (Proposition 8.3.18): if d_S, d_T are associated to generating sets S, T, then there is a positive C such that

$$\frac{1}{C}d_S(x,y) \leq d_T(x,y) \leq Cd_S(x,y).$$

Indeed, every element in S can be represented as a product of elements of T, and vice versa—then we can choose for C the maximum length of such representations.

Therefore, by Theorem 8.4.16, if Γ is hyperbolic with respect to some word metric, then it is hyperbolic with respect to every word metric, and we can speak of a hyperbolic group without specifying a metric. In other words, to be hyperbolic is an algebraic property of a group, and it does not depend on which metric we choose.

The first and dull example of a hyperbolic group is $\mathbb{Z}$. Of course, if Γ is hyperbolic, then so is $\Gamma \times G$ for a finite group G (just because the Gromov-Hausdorff distance between Γ and $\Gamma \times G$ is finite—check this!) Hence $\mathbb{Z} \times G$ is also hyperbolic for any finite G. These examples, however, lack the most interesting features possessed by all other hyperbolic groups. However, already a free (nonabelian) group F_k with $k \geq 2$ generators is a characteristic representative of the hyperbolic world.

Exercise 8.4.26. Prove that a tree is a 0-hyperbolic space, and hence F_k is a hyperbolic group.

Every hyperbolic group "on the large scale looks like a tree" (see Proposition 8.4.14). This property happens to be characteristic for hyperbolic groups, and the latter can also be defined as follows:

Definition 8.4.27. A finitely-generated group Γ is said to be hyperbolic if every subcone at infinity for $C(\Gamma)$ is a topological tree.

It is clear from Proposition 8.4.14 that this definition follows from Definition 8.4.25; we omit the proof of the converse in our exposition. In a sense, hyperbolic groups resemble free groups: when one looks at a hyperbolic groups "from far away", its relations "cannot be seen on the large scale". One can speculate that a "randomly chosen" group has very

high probability to be hyperbolic (for a large number of generators and polynomially many random relations).

A huge source of examples of hyperbolic groups is given by the following proposition:

Theorem 8.4.28. *Let M be a compact Riemannian manifold with sectional curvature bounded above by a negative number (more generally, a compact Alexandrov space of curvature bounded above by $-k$, $k > 0$). Then the fundamental group $\pi_1(M)$ is hyperbolic.*

Proof. The Globalization Theorem for nonpositively curved spaces (Theorem 9.2.9 in Chapter 9) implies that the universal covering space $\tilde{M}$ of M (with the metric lifted from M) has curvature $\leq -k$ in the large. Then by Exercise 8.4.8, $\tilde{M}$ is Gromov hyperbolic. By Corollary 8.3.20, the fundamental group $\pi_1(M)$ is quasi-isometric to $\tilde{M}$. Then Corollary 8.4.17 implies that $\pi_1(M)$ is also Gromov hyperbolic. □

Linear isoperimetric inequality. It turns out that hyperbolic groups possess an unexpected property, which is also characteristic for this class of groups. This property also generalizes a feature of negatively curved manifolds, and we motivate our definitions by the following proposition:

Proposition 8.4.29. *Let M be a simply-connected Riemannian manifold whose sectional curvature is bounded from above by $-k$, $k > 0$. Then there is a number $C = C(k)$ such that, for every closed curve α of length $L > 1$, there exists an immersed disc β whose Riemannian surface area is at most CL and whose boundary is α: $\partial\beta = \alpha$. More generally, if M is a simply-connected length space of curvature bounded above by $-k$, $k > 0$, then there is a number $C = C(k)$ such that every map of a circle to M with length $L > 1$ can be extended to the disc enclosed by the circle so that the 2-dimensional Hausdorff measure of the image is at most CL.*

Note that, in contrast with this inequality, a Euclidean circle of length L cannot be filled by a surface whose area is less than $\frac{1}{4\pi}L^2$: this is the area of the flat disc bounded by the circle.

We leave the proof of the proposition to the reader as an exercise. Here is the idea of an argument. It is based on the following easy observation. Consider a curve of length 1 in the hyperbolic plane of curvature k, and choose a point p. Shortest paths connecting p with points of the curve form a "cone-shaped" region. Then the area of this region is bounded by a number independent of the choice of p (the reader used to "hyperbolic" intuition should not be surprised that while one moves p arbitrarily far away from the curve, the area of this region does not increase to infinity). Indeed, if the

curve is a segment, the "cone-shaped" region is a triangle, and we remember that the area of a triangle cannot exceed π/k by the Gauss-Bonnet formula. Now the reader easily obtains an upper bound for an arbitrary curve by enclosing it into a polygon (fixed for each value of k).

Now to prove the proposition one can choose a point p in M and again consider a cone with its vertex at p by connecting it with all points of the curve by shortest paths. This gives us a topological disc filling in the curve. To estimate its area, one can compare it with a cone over an appropriately chosen curve of the same length in the hyperbolic plane of curvature k (this curve will not be closed any more, but it would have the same distance function to some point as α to p, both with respect to their natural parameters).

Let us formulate a "discrete analog" of the linear isoperimetric inequality given in the proposition. Let Γ be a group with a symmetrical generating set $G = \{g_1, g_1^{-1}, \dots, g_k, g_k^{-1}\}$ and relations $R_1, \dots, R_l$ (where each R_i is a word in letters $g_1, g_1^{-1}, \dots, g_k, g_k^{-1}$ whose product in Γ is identity). Let $\omega = w_1 w_2 \dots w_n$ be a word in letters $g_1, g_1^{-1}, \dots, g_k, g_k^{-1}$. As usual, by its value we mean the product $w_1 \cdot w_2 \cdot \dots \cdot w_n$ in Γ. The value of the empty word is set to be identity.

By a *simple modification* of ω we mean one of the following three operations: inserting one of the words $R_1, \dots, R_l$ anywhere in ω, crossing out a subword of ω identical to one of the words $R_1, \dots, R_l$, or crossing out a generator and its inverse (g_i and g_i^{-1}) if they appear next to each other in ω. Of course, simple modifications do not change the value of ω. It is a standard fact in the theory of finitely-presented groups that two words have the same value if and only if one of them can be transformed into the other by a sequence of simple modifications (the reader can prove this as an easy exercise).

Definition 8.4.30. Γ is said to satisfy a linear isoperimetric inequality if there is a constant C such that every word $\omega = w_1 w_2 \dots w_n$ whose value is the identity can be transformed into an empty word in at most Cn simple modifications.

It turns out that every hyperbolic group satisfies a linear isoperimetric inequality. It is more difficult to show that this property actually characterizes hyperbolic groups. We formulate this as the following alternative definition of hyperbolic groups:

Definition 8.4.31. A finitely-presented group is hyperbolic if it satisfies a linear isoperimetric inequality.

We will not prove here that this definition is equivalent to the two previous definitions (while showing that hyperbolic groups satisfy a linear

isoperimetric inequality is not difficult and can be suggested to the reader as an exercise, a proof of the converse is somewhat involved). Let us, however, discuss isoperimetric inequality for groups in a more geometrical context. Recall that every finitely presented group is the fundamental group of a certain finite two-dimensional cell complex. Indeed, begin with the bouquet of k circles. Choose an orientation for each of the circles and label the circles by pairs of generators (g_i, g_i^{-1}). Then glue a two-cell for each relation following the word representing this relation. This means that the boundary circle of the cell is glued to a curve traversing the circles in the bouquet in the same order as their labels follow in the word representing the relation (and choosing in which direction we follow a circle labeled by (g_i, g_i^{-1}) depending on whether we met g_i or its inverse). The Van Kampen theorem tells us that the fundamental group of this cell complex K is Γ. If each circle comes with an intrinsic metric, and each two-cell is represented by a polygon glued along isometries of its sides, K turns into a length space. Its universal cover $\tilde{K}$ is quasi-isometric to Γ, and hence it is Gromov hyperbolic (note that showing that $K\tilde{K}$ is Gromov hyperbolic, for instance by introducing a metric of negative curvature on $\tilde{K}$, is in its turn a method of arguing that Γ is hyperbolic; see Exercise 8.4.32).

Let us fix a vertex (1-cell) p in $\tilde{K}$. Then a word ω in generators $g_1, g_1^{-1}, \dots, g_k, g_k^{-1}$ determines a path γ in $\tilde{K}$ starting from p and following the edges of the 1-dimensional skeleton of $\tilde{K}$ with labels (lifted from K) corresponding to letters in ω. The value of ω is the identity if and only if this path γ is closed (returns back to p).

What is the meaning of a simple modification of ω in terms of γ? Removing a subword identical to one of the relations means contracting a loop of γ going around one 2-cell; inserting a relation means adding such a loop; finally, crossing out a generator and its inverse next to each other corresponds to contracting a trivial loop (when γ follows an edge and then immediately comes back). Therefore a sequence of simple modifications can be realized by a homotopy of γ, and the number of 2-cells swept by γ in the course of this homotopy is no more than the number of simple modifications in this sequence. Recall that contracting a closed curve to a point sweeps a topological disc. Now if γ is closed (that is, the value of ω is the identity), the number of simple modifications required to transform ω into an empty word is the same as the combinatorial area (the number of 2-cells) in a topological disc bounded by γ. Hence we see that a linear isoperimetric inequality for Γ exactly corresponds to a geometrical linear isoperimetric inequality (like the one we began this section with) for $\tilde{K}$.

Exercise 8.4.32. Show that a group defined by six generators $a_1, a_2, \ldots, a_6$ (here the inverse generators are not listed) and one relation $a_1a_2a_3a_4a_5a_6 = e$ is hyperbolic.

Hint: Construct $\tilde{K}$ as in the discussion above. Note that geometrically it is just a collection of hexagons glued along their sides. Choose the hexagons to be copies of a regular hyperbolic hexagon. Assuming that the hexagon is small (and hence its angles are close to $\frac{2}{3}\pi$) argue that $\tilde{K}$ has Alexandrov curvature bounded from above by -1.

Exercise 8.4.33. Use a direct argument to show that the fundamental group of a negatively-curved compact manifold satisfies a linear isoperimetric inequality.

Hint: This can be done by a refinement of the argument suggested for the proof of Proposition 8.4.29 by representing the manifold as a cell complex (and looking at its 2-dimensional skeleton embedded into the universal cover of the manifold).

We conclude this section with a few remarks, which can perhaps ignite the reader's curiosity and push her to systematic study of this subject. It is well known that the word problem (determining whether the value of a given word is the identity) is algorithmically undecidable for a general finitely presented group. Using a linear isoperimetric inequality, it is easy to see that this problem is always decidable for hyperbolic groups; moreover, a more delicate analysis involving the Morse Lemma shows that it can always be decided in linear time. As a matter of fact, hyperbolic groups belong to the class of *automatic groups*, the groups whose multiplication can be checked by a finite automaton. The reader familiar with this notion can try to re-prove this (elementary, however tricky and very elegant) result of W. Thurston.

8.5. Periodic Metrics

8.5.1. Asymptotic cones of abelian groups. Let G be a finitely generated abelian group equipped with an invariant metric d. We will write $d(g)$ instead of $d(0, g)$ for $g \in G$. It is known from the group theory that G can be decomposed into a sum $\mathbb{Z}^N \oplus G_0$ where $N \geq 0$ is the *rank* of G and G_0 is some finite group (actually the set of finite-order elements of G, called *the torsion*). Then G is at finite Hausdorff distance from its $\mathbb{Z}^N$ component. Such spaces are equivalent for the purposes of the large-scale geometry, which we are concerned with this section. Thus we simply assume that $G \simeq \mathbb{Z}^N$. Then G can be represented as a lattice in an N-dimensional vector space V. The reader may keep in mind the picture of $\mathbb{Z}^N$ as the set

of integer points in $\mathbb{R}^N$, but remember that the Euclidean structure is irrelevant. (In formal algebraic language, the ambient vector space V is obtained from G by tensor multiplication, $V = G \otimes \mathbb{R}$.)

Proposition 8.5.1. *For any $v \in G$, the limit $\lim_{n\to\infty} \frac{d(nv)}{n}$ exists (n is assumed natural).*

Proof. The statement follows from the following lemma, which is a baby version of the subadditive ergodic theorem:

Lemma 8.5.2. *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of nonnegative real numbers such that $x_{m+n} \le x_m + x_n$ for all m, n. Then the limit $\lim_{n\to\infty} \frac{1}{n}x_n$ exists.*

Proof of the lemma. Define $\alpha = \inf_n \frac{x_n}{n}$. We will show that $x_n/n \to \alpha$ as $n \to \infty$. Fix an $\varepsilon > 0$. Then there exist an m such that $x_m/m < \alpha + \varepsilon$. Every natural number n can be represented in the form $n = km + r$ where k is a nonnegative integer and $0 \le r < m$. From the assumptions of the lemma, we have $x_{kn} \le kx_n$ for all k, n (by induction). Hence $x_n \le kx_m + x_r \le km(\alpha+\varepsilon) + x_r \le n(\alpha+\varepsilon) + C(m)$ where $C(m) = \max_{0\le r<m} x_r$. Hence $\limsup_{n\to\infty} x_n/n \le \alpha + \varepsilon$. Since ε is arbitrary, we conclude that $\limsup_{n\to\infty} x_n/n \le \alpha = \inf_n(x_n/n)$, and the lemma follows. □

To derive the proposition from the lemma, let $x_n = d(nv)$. The condition $x_{m+n} \le x_m + x_n$ follows from the triangle inequality. □

The following proposition expresses an important property of the function $v \mapsto \lim_{n\to\infty} \frac{d(nv)}{n}$. Namely, this function can be extended to a semi-norm on the ambient vector space V.

Proposition 8.5.3. *There exists a unique semi-norm $\|\cdot\|$ on V such that $\|v\| = \lim_{n\to\infty} \frac{d(nv)}{n}$ for all $v \in G$.*

Proof. We may assume that $G = \mathbb{Z}^N$ and $V = \mathbb{R}^N$. Let $|\cdot|_E$ denote the Euclidean norm in $\mathbb{R}^N$ and $\{e_i\}_{i=1}^N$ be the standard set of generators. The formula $\|v\| = \lim_{n\to\infty} \frac{d(nv)}{n}$ defines a function on $\mathbb{Z}^N$. Moreover this function is positively homogeneous and satisfies the triangle inequality on $\mathbb{Z}^N$. We can extend this function to $\mathbb{Q}^N$ defining $\|v/n\| = \|v\|/n$ for all $v \in \mathbb{Z}^N$, $n \in \mathbb{N}$. The extended function is also positive homogeneous and satisfies the triangle inequality. Moreover it is a Lipschitz function on $\mathbb{Q}^N \subset \mathbb{R}^N$ (with respect to the Euclidean norm). Indeed, for an $x \in \mathbb{Q}^N$, $x = \sum x_i e_i$, we have $\|x\| \le \sum x_i \|e_i\| \le N^2 \max_i\{\|e_i\|\} \cdot |x|_E$. Since Q^N is dense in $\mathbb{R}^N$, this function has a unique continuous extension to $\mathbb{R}^N$. Positive homogeneity and the triangle inequality remain true by continuity. Hence $\|\cdot\|$ extended to $\mathbb{R}^N$ is a semi-norm. □

The semi-norm $\|\cdot\|$ from the above proposition is called the *asymptotic norm*, or the *stable norm* of the metric d. In all interesting cases, the stable is really a norm (i.e., is strictly positive on nonzero vectors). For example, let d be an orbit metric of a co-compact free totally discontinuous action of G on a length space. Then by Theorem 8.3.19 d is bounded below by the standard Euclidean norm multiplied by some positive constant. This bound is inherited by the stable norm; thus in this case the stable norm is indeed a norm.

Theorem 8.5.4. *Let $\|\cdot\|$ be the stable norm of an invariant metric d on $G \simeq \mathbb{Z}^N$. Assume that $\|\cdot\|$ is a norm. Then*

(1) *$d(v)/\|v\| \to 1$ as $\|v\| \to \infty$ uniformly in $v \in G$.*

(2) *The asymptotic cone of (G,d) is isometric to $(V, \|\cdot\|)$.*

Proof. First observe that the first statement implies the second one. Indeed, one has to prove that $(G, \lambda d) \xrightarrow[GH]{} (V, \|\cdot\|)$ as $\lambda \to 0$ in the sense of pointed space convergence (the distinguished point in 0). In fact, the maps $f_\lambda : (G, \lambda d) \to (V, \|\cdot\|)$ given by $f_\lambda(v) = \lambda v$ satisfy the requirements of the definition of pointed space convergence (Definition 8.1.1). The only nontrivial part is that for every $R > 0$ the distortion of f_λ within the ball of radius R goes to zero (as $\lambda \to 0$), and this follows from the first statement of the theorem.

It remains to prove the first statement. It is trivial that $\|v\| \le d(v)$ for all $v \in G$. On the other hand, by Theorem 8.3.19 we have $d \le C\|\cdot\|_E$ for some constant $C > 0$. Fix an $\varepsilon > 0$ and let S be a finite ε-net of rational vectors (i.e., $S \subset \mathbb{Q}G$) in the unit ball of $(V, \|\cdot\|)$. There exists a large natural number M such that $Mv \in G$ and moreover $d(Mv) \le (1+\varepsilon)M\|v\| \le (1+\varepsilon)M$ for all $v \in S$ (cf. Proposition 8.5.1). For every $w \in G$ there exists a $v \in S$ such that $\big\|w/\|w\| - v\big\| < \varepsilon$. Hence $\big\|Mw/\|w\| - Mv\big\| < M\varepsilon$ and

$$d(Mw)/\|w\| \le d(Mv) + CM\varepsilon \le (1+\varepsilon)M + CM\varepsilon$$

or, equivalently, $d(Mw)/\|Mw\| \le 1 + \varepsilon(C+1)$. Note that this is an upper bound on $d/\|\cdot\|$ for all vectors divisible by M in G. There is a constant $C_1 = C_1(M)$ such that every $v \in G$ has a vector divisible by M in C_1-neighborhood of it. It follows easily that $d(v)/\|v\| \le 1 + \varepsilon(C+2)$ for all $v \in G$ outside the ball of some radius $R = R(M, \varepsilon)$. Since ε is arbitrary, the first statement of the theorem follows. □

Remark 8.5.5. Let d be an orbit metric of a co-compact totally discontinuous action of G on a length space X. Recall that the same action defines a variety of orbit metrics depending on the choice of the orbit. However, the difference between any two orbit metrics is bounded by a constant, so the stable norms of all these metrics coincide. Since the Hausdorff distance

between an orbit and the entire space X is finite, $(V, \|\cdot\|)$ is also the asymptotic cone of X. The norm $\|\cdot\|$ is also called the stable norm of X, or of the quotient space X/G (with respect to a given action or covering).

Remark 8.5.6. The first statement of the theorem remains valid without the assumption that $\|\cdot\|$ is a norm. (The proof is essentially the same.) The second statement requires an obvious correction if $\|\cdot\|$ is not a norm. Namely one has to apply the standard factorization to make $(V, \|\cdot\|)$ a metric space. The result is a normed vector space of a smaller dimension.

Exercise 8.5.7. Let d be a word metric on G defined by a (finite) set S of generators. Prove that the unit ball of the stable norm is the convex hull of the set $S \cup (-S)$ in V.

Exercise 8.5.8. Let d be a word metric on G and $\|\cdot\|$ be its stable norm. Prove that there is a constant C such that $\|x\| \le d(x) \le \|x\| + C$ for all $x \in G$.

Remark 8.5.9. The statement of the last exercise implies that the Gromov–Hausdorff distance between (G, d) and its asymptotic cone $(V, \|\cdot\|)$ is finite. This holds not only for word metrics on $G \simeq \mathbb{Z}^N$ but for any length space X admitting a co-compact action of $\mathbb{Z}^N$ by isometries ([**Bur**]). The proof of this fact in full generality is beyond the scope of this book. In the next subsection we prove it in the special case $X = \mathbb{R}^2$ and $G = \mathbb{Z}^2$ (see Corollary 8.5.13).

8.5.2. Periodic metrics on the plane. In this section we consider universal covers of two-dimensional tori. The two-dimensional torus T^2 is represented as the quotient $\mathbb{R}^2/\mathbb{Z}^2$ where $\mathbb{Z}^2$ acts on $\mathbb{R}^2$ by parallel translations. The projection $p : \mathbb{R}^2 \to \mathbb{R}^2/\mathbb{Z}^2 = T^2$ is a covering map. As we have seen in the previous sections, it defines a 1-1 correspondence between length metrics on T^2 and length metrics on $\mathbb{R}^2$ that are invariant under the action of $\mathbb{Z}^2$. We call the latter class of metrics $\mathbb{Z}^2$-*periodic* ones. A typical example is a Riemannian metric on $\mathbb{R}^2$ defined by a bi-periodic metric tensor (the reader may visualize it by thinking of a "wavy surface" whose bumps are arranged periodically by repeating the same pattern). We are going to see that the large-scale geometry of such a surface is the same as of a certain two-dimensional normed space; moreover, this surface cannot be distinguished from the space by distance measurements whose error exceeds some number (notice that we mean the *absolute* precision, as opposed to the relative one). This conclusion is actually true in any dimension, although we will use some essentially two-dimensional arguments for the sake of simplicity.

The fundamental group of T^2 is $\mathbb{Z}^2$. If $\gamma : [a, b] \to T^2$ is a closed curve (loop) and $\tilde\gamma$ is a lift of γ in $\mathbb{R}^2$, then the vector $[\gamma] := \tilde\gamma(b) - \tilde\gamma(a)$ is the element of $\mathbb{Z}^2$ represented by γ. (Note that this vector does not depend on

the choice of $\tilde{\gamma}$.) It is called the *rotation vector* of γ. Two loops have equal rotation vectors if and only if they belong to the same free homotopy class.

Let T^2 be equipped with a length metric d and let $\tilde{d}$ denote the corresponding $\mathbb{Z}^2$-periodic metric on $\mathbb{R}^2$. For every $v \in \mathbb{Z}^n$ we denote by $\ell(v)$ the length of a shortest loop in T^2 whose rotation vector is v. In terms of the metric $\tilde{d}$, this can be written as follows:

$$\ell(v) = \min_{x \in \mathbb{R}^2} \tilde{d}(x, x+v).$$

Indeed, $\tilde{d}(x, x+v)$ equals the length of a shortest path from x to $x+v$ whose projection to T^2 is a shortest loop with rotation vector v and endpoints at $p(x)$.

The function $x \mapsto \tilde{d}(x, x+v)$ is a continuous bi-periodic function on $\mathbb{R}^2$; hence it attains its maximum and minimum. This validates the right-hand side of the formula for $\ell(v)$ and also implies that there is a constant C such that $\tilde{d}(x, x+v) \le \ell(v) + C$ for all $x \in \mathbb{R}^2$.

Let $\|\cdot\|$ be the stable norm of $(\mathbb{R}^2, \tilde{d})$. Theorem 8.5.4 can be reformulated as follows:

$$\lim_{\|x-y\| \to \infty} \frac{\tilde{d}(x,y)}{\|x-y\|} = 1 \qquad (x, y \in \mathbb{R}^2).$$

The definition of $\|\cdot\|$ and the estimate $\tilde{d}(x, x+v) \le \ell(v) + C$ implies that

$$\|v\| = \lim_{n \to \infty} \frac{\tilde{d}(x, x+nv)}{n} = \lim_{n \to \infty} \frac{\ell(nv)}{n}$$

for a $v \in \mathbb{Z}^2$. One of the purposes of this section is to improve these formulas.

The above considerations could apply to tori and periodic metrics in any dimension. The following theorem is valid only in dimension 2.

Theorem 8.5.10. *For every $v \in \mathbb{Z}^2$ and $n \in \mathbb{N}$, one has $\ell(nv) = n\ell(v)$.*

Remark 8.5.11. The theorem can be interpreted as follows: if γ is a minimal loop in its free homotopy class in T^2, then γ^n (the same loop passed around n times) is also minimal in its free homotopy class. Note that the lift of a minimal loop is a shortest path in $(\mathbb{R}^2, \tilde{d})$. Suppose that γ is parameterized by arc length, $\gamma : [0, L] \to T^2$. Let $\gamma_0 : \mathbb{R} \to T^2$ be an L-periodic curve (i.e., such that $\gamma_0(t+L) = \gamma_0(t)$ for all t) whose restriction on $[0, L]$ is γ. Then a lift $\tilde{\gamma}_0$ of γ_0 is a geodesic line—every interval of $\tilde{\gamma}_0$ is a shortest path.

Before proving the theorem, we formulate two immediate corollaries. As we already mentioned, Corollary 8.5.13 can be generalized to any dimension.

Corollary 8.5.12. *$\|v\| = \ell(v)$ for all $v \in \mathbb{Z}^2$.*

Corollary 8.5.13. *There is a constant C such that $\left|\tilde{d}(x,y) - \|x-y\|\right| \le C$ for all $x, y \in \mathbb{R}^2$.*

Proof. Under an additional condition that $x - y \in \mathbb{Z}^2$, the existence of an upper bound for $\left|\tilde{d}(x,y) - \|x-y\|\right|$ follows from the previous corollary. Since $\mathbb{R}^2/\mathbb{Z}^2$ is compact, the statement follows for all $x, y \in \mathbb{R}^2$. □

Proof of Theorem 8.5.10. We will need the following topological lemma

Lemma 8.5.14. *Let $v \in \mathbb{Z}^2$ be a nonzero vector and γ be a loop in T^2 with rotation vector nv where $n \in \mathbb{N}$, $n > 1$. Then γ is not simple, i.e., has self-intersections. Moreover, γ can be decomposed into two loops (with endpoints at one of the self-intersections) whose rotation vectors equal v and $(n-1)v$.*

Proof. This fact is well-known in elementary topology. Here we outline one of the possible proofs.

Let A denote the quotient of $\mathbb{R}^2$ by the group of integer multiples of nv. This quotient is homeomorphic to the cylinder $S^1 \times \mathbb{R}$ so that the translation by v in A corresponds to the rotation by $2\pi/n$. We denote the rotation by R. There is a natural covering map from A to T^2 that commutes with the two projections from $\mathbb{R}^2$ to A and T^2. Let $\tilde{\gamma}$ be a lift of γ in A. Then $\tilde{\gamma}$ is a closed noncontractible curve. Such a curve separates the cylinder's "ends" from each other; i.e., points of $S^1 \times \mathbb{R}$ with the last coordinate near $+\infty$ and those with the last coordinate near $-\infty$ belong to different components of the curve's complement. Consider the curve $R(\tilde{\gamma})$ in $S^1 \times \mathbb{R}$. We will show that it intersects γ at some point. Suppose the contrary; then $R(\tilde{\gamma})$ is contained in one connected component of the complement of $\tilde{\gamma}$. Let x^+ and x^- be two points on (the image of) $\tilde{\gamma}$ in $S^1 \times \mathbb{R}$ where the last coordinate attains its maximum and minimum, respectively. The points $R(x^+)$ and $R(x^-)$ are on the curve $R(\tilde{\gamma})$. On the other hand, $R(x^+)$ and $R(x^-)$ belong to different components of the complement of $\tilde{\gamma}$ because they can be connected to the $+\infty$ and $-\infty$ ends, respectively. This is a contradiction.

Thus $R(\tilde{\gamma})$ intersects $\tilde{\gamma}$. In other words, there is a point x in $\tilde{\gamma}$ such that $y = R(x)$ also belongs in $\tilde{\gamma}$. These two points split $\tilde{\gamma}$ into two arcs whose projections to T^2 are closed loops with rotation vectors $\pm v$ and $(n \mp 1)v$. If these rotation vectors are v and $(n-1)v$, the lemma follows. Otherwise, we can apply the same argument to the loop with the rotation vector $(n+1)v$, and iterate this until we obtain a sub-loop with rotation vector v (not $-v$). This must eventually happen because otherwise γ have sub-loops with arbitrary large rotation vectors and this contradicts the compactness of its lift in $\mathbb{R}^2$. □

Now the theorem follows by induction in n. It is trivial for $n = 1$. If $n > 1$, let γ be a minimal loop representing the vector nv. By the lemma, we can decompose γ into two loops γ_1 and γ_2 with rotation vectors v and $(n-1)v$ respectively. Then $\ell(nv) = L(\gamma) = L(\gamma_1)+L(\gamma_2) \geq \ell(v)+\ell((n-1)v) = n\ell(v)$ using induction. The inverse inequality $\ell(nv) \leq n\ell(v)$ is trivial. □

Exercise 8.5.15. Let d be a smooth Riemannian $\mathbb{Z}^2$-periodic metric on $\mathbb{R}^2$. Prove that the unit ball of its stable norm $\|\cdot\|$ is strictly convex in the sense that its boundary does not contain straight line segments. Equivalently, $\|v + w\| < \|v\| + \|w\|$ whenever v and w are linearly independent.

Hint: Any two loops in T^2 with linearly independent rotation vectors have an intersection point.

Exercise 8.5.16. Let ℓ_1, ℓ_2 and ℓ_3 be straight lines in R^3 that are parallel to the three coordinate axes, intersect the open unit cube $(0,1)^n$, and do not intersect one another. Prove that there is a $\mathbb{Z}^3$-periodic Riemannian metric in $\mathbb{R}^3$ such that

(1) ℓ_1, ℓ_2 and ℓ_3 are minimal geodesics.

(2) There are no other minimal geodesics except these lines and their integer parallel translations.

(3) The stable norm of the metric is $\|\cdot\|_1$, where $\|(x,y,z)\| = |x| + |y| + |z|$.

Hint: Let the metric tensor be standard Euclidean at these lines and sufficiently large outside appropriate neighborhoods of them. (This example and related topics are discussed in [**Ban**].)

8.5.3. Asymptotic volumes of periodic metrics. Consider a $\mathbb{Z}^n$-periodic metric d in $\mathbb{R}^n$. This metric, or its restriction to $\mathbb{Z}^n$, defines the stable norm $\|\cdot\|$ in $\mathbb{R}^n$, and Theorem 8.5.4 implies that the quantity $d(x,y)/\|x - y\|$ tends to 1 as $\|x - y\|$ or $d(x,y)$ go to infinity. In other words, the metric d "at large scale" is equivalent to the distance function defined by $\|\cdot\|$. (In fact, their difference is bounded above by a constant; see Remark 8.5.9.) We denote the unit ball of the stable norm by D, i.e., $D = \{v \in \mathbb{R}^n : \|v\| \leq 1\}$.

We will study the volumes of balls of large radii in the metric d.

Definition 8.5.17. Let d be as above. Its *asymptotic volume*, denoted $\Omega(d)$, is defined by

$$\Omega(d) = \lim_{r\to\infty} \frac{\mathrm{Vol}_d(B_r(x))}{r^n}$$

where $x \in \mathbb{R}^n$ is an arbitrary point and $B_r(x)$ is the ball of radius r in the metric d.

The first thing to prove is that the limit in the definition exists.

Proposition 8.5.18. *The asymptotic volume $\Omega(d)$ exists and does not depend on $x \in \mathbb{R}^n$. Moreover, in the notation introduced above*

$$\Omega(d) = \mathrm{Vol}_d(I^n) \cdot \mu_n(D)$$

where $I^n = [0,1]^n$ and μ_n is the standard Lebesgue (or Hausdorff) measure in $\mathbb{R}^n$. In particular, $\Omega(d)$ is finite and positive.

Proof. We give a sketch of a proof leaving the details to the reader. Theorem 8.5.4 implies that $d(x,y) \sim \|x-y\|$ as $\|x-y\| \to \infty$. Hence a ball $B_x(r)$ of d is close to the set $rD = \{rv : v \in D\}$ (i.e., the r-ball of the norm $\|\cdot\|$) in the following sense: for every given $\varepsilon > 0$, $(1-\varepsilon)rD \subset B_x(r) \subset (1+\varepsilon)D$ for all sufficiently large r. This allows us to replace $B_r(x)$ by rD in the definition of the asymptotic volume. The space $\mathbb{R}^n$ is split into cells obtained by translating the cube I^n by all integer vectors. All their volumes are equal because the metric is periodic. Let $Q(r)$ denote the union of cells contained entirely in rD. Then $\mathrm{Vol}_g(rD) \sim \mathrm{Vol}_g(Q(r))$ and $\mu_n(rD) \sim \mu_n(Q(r))$ as $r \to \infty$ because the number of cells that intersect the boundary of rD is negligible. Since $\mathrm{Vol}_g(Q(r))/\mu_n(Q(r)) = \mathrm{Vol}_g(I^n)/\mu_n(I^n)$, it follows that

$$\mathrm{Vol}_g(B_r(x)) \sim \frac{\mathrm{Vol}_g(I^n)}{\mu_n(I^n)}\mu_n(rD) = \frac{\mathrm{Vol}_g(I^n)}{\mu_n(I^n)}\mu_n(D)r^n, \qquad r \to \infty.$$

The proposition follows. □

Exercise 8.5.19. Prove the above proposition for any Riemannian manifold M whose metric is invariant under a proper co-compact action of $\mathbb{Z}^n$. The formula for $\Omega(d)$ should be changed as follows: replace $\mathrm{Vol}_d(I^n)$ by $\mathrm{Vol}_{\bar{d}}(M/\mathbb{Z}^n)$ where $(M/\mathbb{Z}^n, \bar{d})$ is the quotient space of the action.

Our purpose is to prove that the asymptotic volume can be bounded below by a constant depending only on dimension but not on the metric. In fact, the minimal possible asymptotic volume equals the volume of the Euclidean unit ball and is achieved for flat Euclidean metrics (this is proved in [**BI**]). The lower bound given in the following theorem is not optimal; however the mere existence of a common lower bound for all metrics is not obvious.

Theorem 8.5.20. $\Omega(d) \geq (2/n)^n$ *for any $\mathbb{Z}^n$-periodic Riemannian metric d in $\mathbb{R}^n$.*

Proof. Let $\|\cdot\|$ be the stable norm of d, D its unit ball and Q be an affine cube such that $\frac{1}{n}Q \subset D \subset Q$ (such a Q exists due to Lemma 5.5.19). Since $D \subset Q$, one has $\|v - v'\| \geq 2$ whenever v and v' belong to opposite faces

of Q. The idea of the proof is the following: we let $x = 0$ replace the ball $B_r(x) \approx rD$ in the definition of $\Omega(d)$ by a smaller set $\frac{r}{n}Q$. Then the volume of $\frac{r}{n}Q$ is estimated below by means of the Besikovitch inequality.

Now we pass to a formal argument. Fix an $\varepsilon > 0$ and let r be so large that

$$(1-\varepsilon)\|x-y\| \le d(x,y) \le (1+\varepsilon)\|x-y\|$$

whenever $\|x-y\| \ge r/n$ or $d(x,y) \ge r/n$ (see Theorem 8.5.4 and the beginning of this section). Then the distance (in d) between the opposite faces of $\frac{r}{n}Q$ is not less than $2(1-\varepsilon)r/n$. By the Besikovitch inequality it follows that

$$\mathrm{Vol}_d\left(\frac{r}{n}Q\right) \ge (1-\varepsilon)^n(2r/n)^n.$$

On the other hand, since $\frac{r}{n}Q \subset rD$, we have $\|x\| \le r$ and hence $d(0,x) \le (1+\varepsilon)r$ for all $x \in \frac{r}{n}Q$. This means that our affine cube $\frac{r}{n}Q$ is contained in the ball $B_{(1+\varepsilon)r}(0)$. Thus

$$\Omega(d) = \lim_{r\to\infty} \frac{\mathrm{Vol}_d(B_{(1+\varepsilon)r}(0))}{(1+\varepsilon)^n r^n} \ge \liminf_{r\to\infty} \frac{\mathrm{Vol}_d(\frac{r}{n}Q)}{(1+\varepsilon)^n r^n} \ge (2(1-\varepsilon)/n)^n/(1+\varepsilon)^n.$$

Since ε is arbitrary, the desired inequality $\Omega(d) \ge (2/n)^n$ follows. □

The following exercise improves the constant $(2/n)^n$ in the theorem.

Exercise 8.5.21. For a unit ball D of a norm in $\mathbb{R}^n$ define $\lambda(D)$ by

$$\lambda(D) = \sup\{\frac{\mu_n(D)}{\mu_n(Q)} : Q \text{ is an affine cube containing } D\}.$$

Prove that $\Omega(d) \ge 2^n\lambda(D)$ for any periodic Riemannian metric in $\mathbb{R}^n$ where D is the unit ball of the stable norm.

Hint: Combine the proofs of Proposition 8.5.18 and Theorem 8.5.20. Namely, show that

1. $\mathrm{Vol}_d(rQ) \sim r^n \mathrm{Vol}_d(I^n) \cdot \mu_n(Q)$ as $r \to \infty$ for any affine cube Q;

2. $\mathrm{Vol}_d(rQ) \ge (2r)^n - o(r^n)$ as $r \to \infty$ if an affine cube Q contains the unit ball of the stable norm.

Exercise 8.5.22. Prove that for every integer $n \ge 2$ and every $C > 0$ there exists a periodic Riemannian metric d in $\mathbb{R}^n$ with $\Omega(d) \ge C$ (compare with Exercise 5.6.16).

Chapter 9

Spaces of Curvature Bounded Above

In this chapter we concentrate on properties specific for spaces with upper curvature bounds. In Section 9.1 we deal with local properties of such spaces, and Section 9.2 contains an introduction to the global geometry of Hadamard spaces (that is, complete simply connected spaces of nonpositive curvature). We will see that the local properties of spaces of curvature $\le k$ almost do not depend on k. However, the global properties can be very different for the cases $k \le 0$ and $k > 0$. Actually not much is known about global structure of spaces with arbitrary (possibly positive) curvature bounds, even if the space is a Riemannian manifold or a surface in $\mathbb{R}^3$. One of the reasons is that these classes of spaces are defined in terms of local properties which may not hold "in the large" (see e.g. Example 9.1.13.) In particular, these classes are not stable w.r.t. Gromov–Hausdorff convergence.

In the case $k \le 0$ the situation is completely different: for simply connected spaces the local curvature conditions imply that the same distance and angle comparison properties hold for all (arbitrarily large) triangles. This is the contents of Globalization Theorem 9.2.9. This theorem allows us to obtain much more information about the global geometry than we have in the general case.

For readers familiar with Riemannian geometry we point out that almost all classical theorems on Riemannian manifolds of nonpositive (and strictly negative) curvature have generalizations for general spaces of curvature ≤ 0 (resp. $\le k < 0$). These are theorems about the fundamental group, group of isometries, existence of flat planes and asymptotic behavior at infinity (ideal boundary). The proofs are very similar to ones in Riemannian geometry,

and this indicates that the differential technique in the Riemannian proofs is not that essential. On the contrary, convexity arguments often play the key role.

9.1. Definitions and Local Properties

9.1.1. Definitions. We begin with the definition of a space of curvature $\le k$, where k is an arbitrary real number. All components of this definition can be found somewhere in Chapter 4; here we just collect them together.

Recall that the *k-plane* is the two-dimensional model space of constant curvature k. Depending on the sign of k, this is either the Euclidean plane, a sphere (of radius $\sqrt{k}$) with its length metric, or a hyperbolic plane of curvature k. We denote by R_k the diameter of the k-plane, i.e., $R_k = 1/\sqrt{k}$ if $k > 0$ and $R_k = \infty$ if $k \le 0$.

A *comparison triangle* for a triangle $\triangle abc$ in a length space is a triangle $\triangle \bar{a}\bar{b}\bar{c}$ in the k-plane such that $|\bar{a}\bar{b}| = |ab|$, $|\bar{a}\bar{c}| = |ac|$ and $|\bar{b}\bar{c}| = |bc|$. When talking of comparison triangles, we always assume that the triangle's perimeter $|ab| + |bc| + |ac|$ is less than $2R_k$ (of course, this imposes no restriction if $k \le 0$). This assumption guarantees that a k-comparison triangle $\triangle \bar{a}\bar{b}\bar{c}$ exists and is unique up to an isometry.

Definition 9.1.1. A *space of curvature* $\le k$ is a length space X which can be covered by a family of open sets $\{U_i\}_{i\in I}$ so that every U_i satisfies the following:

1. Every two points in $U - I$ can be connected by a shortest path in U_i.

2. For any $a, b, c \in U_i$ such that $|ab| + |bc| + |ac| < 2R_k$ and a point d in any shortest path $[ac]$, the "triangle condition" holds: $|db| \le |\bar{d}\bar{b}|$, where $\triangle \bar{a}\bar{b}\bar{c}$ is a comparison triangle for $\triangle abc$ in the k-plane and $\bar{d}$ is the point in $[\bar{a}\bar{c}]$ such that $|ad| = |\bar{a}\bar{d}|$.

A *space of curvature bounded above* is a length space in which every point has a neighborhood where these two conditions are satisfied for some k (possibly varying from one point to another). Equivalently, every point has a neighborhood whose induced length metric is of curvature $\le k$ for some k depending on a point.

In fact, it is sufficient to verify the triangle condition only for d being a midpoint between a and c (Exercise 4.1.11). Note that the definition does not require the uniqueness of a shortest path $[a, c]$ (or the uniqueness of a midpoint). However this uniqueness follows automatically as we will see in Subsection 9.1.3.

Definition 9.1.1 is a modification of Definition 4.1.9. Other definitions of a nonpositively curved space (4.1.2, 4.1.9, and 4.3.1) can be similarly

modified for the case of curvature $\leq k$, and the four resulting definitions are equivalent (Exercise 4.6.3).

An unpleasant part of our definition is the requirement that any two points (in a neighborhood U) can be connected by a shortest path. While we mainly restrict ourselves to locally compact spaces where shortest paths always exist, not locally compact spaces can arise in certain constructions. In order to handle them, it is natural to have a more general version of Definition 9.1.1 where midpoints are replaced by "almost midpoints" (similarly to the "intrinsic versus strictly intrinsic" case in Subsection 2.4.3).

Definition 9.1.2. A *space of curvature* $\leq k$ is a length space that can be covered by a family of open sets $\{U_i\}$ so that each U_i satisfies the following:

For every three points a, b, $c \in U_i$, every $\varepsilon > 0$ and every ε-midpoint d between a and b, the distance $|bd| \leq |\bar{b}\bar{d}| + f(\varepsilon)$, where $\bar{d}$ is the midpoint of the side $[\bar{a}\bar{c}]$ in a comparison triangle $\bar{a}\bar{b}\bar{c}$ in k-plane, and $f(\varepsilon) \to 0$ as $\varepsilon \to 0$ (for a, b, c fixed).

In fact, just Definition 9.1.2 is the "standard" one, but we stick to Definition 9.1.1 to avoid technical complications. Fortunately, the two definitions are equivalent for complete spaces:

Proposition 9.1.3. *For complete (more generally, locally complete) spaces Definitions 9.1.1 and 9.1.2 are equivalent.*

Proof. Suppose that a complete length space X has curvature $\leq k$ in the sense of Definition 9.1.2. We only need to prove the local existence of shortest paths in X. By Theorem 2.4.16 (more precisely, its "localized" version—formulate and prove it yourself) it is sufficient to prove the existence of midpoints. Let $a, b \in U$ where $U = U_i$ is from the formulation. We will construct a midpoint between a and b as a limit of a Cauchy sequence of "almost midpoints".

Let $\{\varepsilon_n\}$ be a sequence of positive numbers, $\varepsilon_n \to 0$, and let d_n be an ε_n-midpoint for a, b. We apply triangle comparison to the triangle $\triangle abd_n$ and ε_m-midpoint d_m. Then $|d_nd_m| \leq |\bar{d}_n\bar{d}| + f(\varepsilon_m)$, where $\bar{d}$ is the midpoint in a comparison triangle $\triangle\bar{a}\bar{d}_n\bar{b}$. It is easy to see that $|\bar{d}_n\bar{d}| \to 0$.

(Indeed, the point $\bar{d}_n$ in the k-plane belongs to the intersection of the balls of radius $|ab|/2 + \varepsilon_n$ centered at $\bar{a}$ and $\bar{b}$. The diameter of this intersection goes to zero since $|ab|$ is fixed and $\varepsilon_n \to 0$.)

Therefore we can choose a subsequence of indices n such that the corresponding points d_n form a Cauchy sequence. Since X is complete, the subsequence converges to a point which is obviously a midpoint between a and b.

Note that this midpoint d is unique: otherwise we could apply the curvature condition to the triple a, b, d and once more midpoint $\tilde{d}$ to come to a contradiction. □

9.1.2. Examples. In Chapter 4 we have already discussed several examples of graphs and polyhedra of nonpositive curvature. Even these first examples show that a nonpositively curved space may have a very complicated structure. For example (unlike in the case of nonnegative curvature), the local dimension of a nonpositively curved space may vary from point to point. (Though the notion of dimension is not defined yet, its meaning is clear in simple examples like two-polyhedra).

Example 9.1.4. An open subset of a space of curvature $\le k$ with its induced length metric is itself a space of curvature $\le k$. In particular, an open subset of $\mathbb{R}^n$ is a nonpositively curved space.

Note that the word "open" is essential. For example, if one removes closed balls $B_{r_k}(1/k)$ with $r_k = 1/(k+1)^2$, $k = 1, 2, \dots$, from the plane, the resulting length space is not a space of nonpositive curvature. See also Example 9.1.7 below.

Example 9.1.5. The direct product of spaces of curvature not greater than k is a space of curvature not greater than $\max\{k, 0\}$.

The proof is straightforward. Compare with Exercise 4.1.13

Note that a product of spaces of curvature $\le k$ is *not* a space of curvature $\le k$ if $k < 0$ unless one of the multiplied spaces is a single point. Indeed, consider two arbitrary shortest paths in spaces X and Y. These shortest paths (as subspaces of X and Y) are isometric to intervals of $\mathbb{R}$; hence their product in $X \times X$ is a convex set isometric to a subset of $\mathbb{R}^2$. Therefore $X \times Y$ cannot have strictly negative curvature.

Example 9.1.6. The complement of an open round disc in the plane $\mathbb{R}^2$ (with its induced length metric) is nonpositively curved. (Exercise: verify this.)

More generally, any locally simply connected subset $X \subset \mathbb{R}^2$ (with its induced length metric) is nonpositively curved. In outline, the proof goes as follows. By the Jordan Curve Theorem, any triangle in X bounds a region homeomorphic to the disc. Local simple-connectedness implies that, if the triangle is small enough, the region that it bounds is contained in X. It then follows that the sides of the triangle are concave arcs (in the usual planar sense) because otherwise one could shorten them in the bounded region. And the angle comparison property for such "concave" triangles is more or less trivial (if one pulls the vertices apart to straighten the sides, angles increase).

The above example is a purely two-dimensional phenomenon. The next one shows that similar constructions in $\mathbb{R}^3$ may fail.

Example 9.1.7. The complement of an open ball in $\mathbb{R}^3$ is *not* a space of nonpositive curvature. For example, consider a triangle whose vertices belong to the boundary of the removed ball. Its sides are shortest paths in the sphere and its angles are *greater* than ones of the comparison triangle in the plane.

Exercise 9.1.8. Is the complement of an open ball in $\mathbb{R}^3$ a space of curvature $\leq k$ for some k?

Answer: yes, for $k = 1$.

Example 9.1.7 shows that the completion of a nonpositively curved space may fail to be nonpositively curved. In the following example the completion of a nonpositively curved space does not have any upper curvature bound at all.

Example 9.1.9. A circular cone in $\mathbb{R}^3$ with its origin removed is a flat space and hence is nonpositively curved. Its completion—a cone with the origin—is not a space of curvature bounded above.

Example 9.1.10. As explained in Chapter 5, a two-dimensional surface (more generally, a Riemannian manifold) is a space of curvature $\leq k$ if and only if its Gauss curvature is less than or equal to k everywhere.

Example 9.1.11 (Space forms)**.** Every orientable closed surface of genus $n \geq 2$ can be equipped with a metric of constant negative Gauss curvature. A surface of genus n can be (topologically) glued from a polygon with $4n$ vertices as follows: mark the (cyclically ordered) sides of a $4n$-gon by the symbols a_1, b_1, a_1^{-1}, b_1^{-1}, $\cdots$, a_n, b_n, a_n^{-1}, b_n^{-1}; then glue together the sides marked by the same letters, say a and a^{-1}. To obtain a surface equipped with a metric, apply this gluing construction to a convex hyperbolic $4n$-gon satisfying the following conditions: the sides glued together have equal lengths, and the sum of the polygon's angles equals 2π.

It is easy to see that such a polygon exists; for example, it can be found among regular polygons. Indeed, if a polygon is small, the sum of its angles approximately equals that in a Euclidean polygon: $(4n - 2)\pi > 2\pi$. And if it is large (i.e., if the vertices go to the ideal boundary), the angles tend to zero. By continuity, there exists a regular $4n$-gon whose sum of angles equals 2π; see also 5.3.5. (In fact, there is a continuum of different polygons suitable for this constructions.)

Gluing sides of such a polygon according to the rule yields an orientable surface of constant curvature -1 which is locally isometric to the hyperbolic

plane (prove this). A constant curvature $k < 0$ can be then obtained by multiplying the metric by $1/\sqrt{-k}$.

Closed surfaces of constant negative curvature are called (two-dimensional) *space forms.*

Exercise 9.1.12. Modify this construction

a) to obtain closed nonorientable surfaces of any genus $n \geq 3$ with a metric of constant negative curvature;

b) to obtain complete noncompact surfaces (with cusps) having finite area.

Example 9.1.13. Take two unit spheres, make a small round hole in each of them, and glue them to a thin cylinder attaching its boundary circles to the boundaries of the holes. (We assume the lengths of the boundary cycles are equal.) This yields a closed surface of curvature ≤ 1 which looks like a dumbbell. It can be approximated by a smooth one if desired. Note that such a surface can contain an arbitrarily short closed geodesic provided that the cylinder is sufficiently thin.

The following nontrivial result allows us to obtain new examples of nonnegatively curved spaces.

Theorem 9.1.14 (S. Alexander, R. Bishop [**AB1**]). *Let X and Y be complete simply connected spaces of nonpositive curvature, and $f\colon X \to \mathbb{R}$ be a concave function. Then the warped product $X \times_f Y$ is a (complete simply connected) space of nonpositive curvature.*

9.1.3. Elementary properties.

Definition 9.1.15. Let X be a space of curvature $\leq k$. A *normal ball* in X is a metric ball U of radius less than $R_k/2$ satisfying the comparison conditions from Definition 9.1.1. (Sometimes we also call such a ball a *normal region* or a *normal neighborhood.*)

For a point $p \in X$, the supremum of numbers r such that $B_r(p)$ is a normal ball is called the *convexity radius* at p and is denoted by $r(p)$. It is easy to prove (do this yourself) that the number $r(K)$ defined as $\inf\{r(p) : p \in K\}$ is positive for each compact set K. The number $r(K)$ is called the *convexity radius* of K.

The term "convexity radius" is motivated by the following

Proposition 9.1.16. *Every normal ball is convex.*

Proof. Let $B_r(p)$ be a normal ball, $a, b \in B_r(p)$. Since $r < R_k/2$, a ball $B^k_r(\bar{p})$ in the k-plane is a convex set. By the comparison property for a

comparison triangle $\triangle\bar{p}\bar{a}\bar{b}$, one has $|\bar{p}\bar{d}| \le r$ for any point $\bar{d} \in [\bar{a}\bar{b}]$. Hence $|pd| \le |\bar{p}\bar{d}| \le r$ for any $d \in [a, b]$, and this means that $[ab] \subset B_r(p)$. □

All the statements in the next proposition follow almost immediately from the definitions.

Proposition 9.1.17. *Let X be a complete locally compact space of curvature $\le k$ and $U \subset X$ be a normal ball. Then*

(1) *For every two points $a, b \in U$ there is a unique shortest path connecting these points, and this path is contained in U.*
(2) *If γ_1, γ_2 are curves in U, then the shortest path $[\gamma_1(t)\gamma_2(s)]$ depends continuously on t, s.*
(3) *Every ball $B_r(p) \subset U$ is convex.*
(4) *Let $[ab], [ac]$ be two shortest paths starting at a. If $\angle bac = \pi$, then the curve bac is a shortest path.*
(5) *Every geodesic contained in U is a shortest path.*

Proof. (1) Suppose there are two shortest paths, $[ab]_1$ and $[ab]_2$. For every point $c \in [ab]_1$ one can consider the triangle $\triangle abc$ and its comparison triangle $\triangle\bar{a}\bar{b}\bar{c}$. The latter one is degenerate: $\bar{c} \in \bar{a}\bar{b}$. Let d be a point of $[ab]_2$ such that $|ac| = |ad|$. Then it has to be $|cd| \le |\bar{c}\bar{d}| = 0$, so $c = d$ and $[ab]_1 = [ab]_2$ since c is arbitrary.

Since a limit of shortest paths is a shortest path, (2) follows from (1).

The proof of (3) is the same as for Proposition 9.1.16. And (4) follows from the fact that a comparison triangle for $\triangle abc$ is degenerate with $\angle\bar{b}\bar{a}\bar{c} = \pi$.

(5) Let $\gamma : [0, L] \to U$ be a geodesic but not a shortest path. Let $t \in [0, L]$ be the maximal value of the parameter such that the restriction of γ to $[0, t]$ is still a shortest path. Such a maximal t exists because the restriction to $[0, t]$ is a shortest path if and only if $|\gamma(0)\gamma(t)| = t$, and the latter condition defines a closed set. Since γ is a geodesic, its restriction to $[t, t + \varepsilon]$ is a shortest path for a sufficiently small $\varepsilon > 0$. Applying the statement (4) to the shortest paths $\gamma|_{[0,t]}$ and $\gamma_{[t,t+\varepsilon]}$ yields a contradiction with the maximality of t. □

Remark 9.1.18. Proposition 9.1.17 implies that a normal ball U is contractible. Indeed, fix a point $p \in U$; then for every $x \in U$ there is a unique shortest path $\gamma_x \colon [0, 1] \to U$ (parameterized with constant speed) such that $\gamma_x(0) = x$ and $\gamma_x(1) = p$. Then a map $H \colon U \times [0, 1] \to U$ defined by $H(x, t) = \gamma_x(t)$ is a homotopy between the identity map id_U and the map which sends all points of U to p. Geometrically, this homotopy moves every

point $x \in U$ towards p along the corresponding shortest path. (Exercise: prove that H is continuous.)

In particular, every space of nonpositive curvature is locally simply connected and therefore has a universal covering space (cf. Theorem 3.4.11).

Simple examples, like flat tori or hyperbolic space forms, show that, in general, the local curvature conditions do not imply analogous properties in the large. This is related to the fact that these spaces contain a closed geodesic. Tori and surfaces of higher genus are not simply connected, so closed geodesics can be found by minimizing the length in a free homotopy class of noncontractible closed curve. The space in Example 9.1.13 is simply connected (it is homeomorphic to the sphere) but its convexity radius can be made arbitrarily small while the curvature bound stays the same. We will see later that this effect (small convexity radius in a simply connected space) is possible only for $k > 0$.

9.1.4. Extreme cases of comparison conditions. The definition of a space of curvature $\le k$ says that distances and angles in (sufficiently small) triangles must satisfy certain inequalities. The following proposition tells us what happens if these inequalities turn into equalities.

Proposition 9.1.19. *Let $\triangle abc$ be a triangle in a normal region of a space X of curvature $\le k$ and $\triangle \bar{a}\bar{b}\bar{c}$ a comparison triangle in the k-plane. Then the following three conditions are equivalent:*

(i) The angle at a of the triangle $\triangle abc$ is equal to the angle at $\bar{a}$ of $\triangle \bar{a}\bar{b}\bar{c}$.

(ii) For some point $x \in [ac]$, $d \ne a, c$, the equality $|bx| = |\bar{b}\bar{x}|$ holds, where $\bar{x} \in [\bar{a}\bar{c}]$ is a point such that $|ax| = |\bar{a}\bar{x}|$.

(iii) All angles of $\triangle abc$ are equal to the corresponding angles of $\triangle \bar{a}\bar{b}\bar{c}$.

(iv) $\triangle abc$ spans a totally geodesic surface isometric to the triangle $\bar{a}\bar{b}\bar{c}$ with its interior (i.e., the region in the k-plane bounded by $\triangle \bar{a}\bar{b}\bar{c}$). More formally, there exists a distance-preserving map from the "full" triangle $\bar{a}\bar{b}\bar{c}$ to X which maps the sides of $\triangle \bar{a}\bar{b}\bar{c}$ to the respective sides of $\triangle abc$.

Proof. Let α, β, γ denote the angles at a, b, c of $\triangle abc$ and $\bar{\alpha}$, $\bar{\beta}$, $\bar{\gamma}$ the respective angles of the comparison triangle.

1. It is obvious that (iv) implies the assertions (i)–(iii) and (iii) implies (i).

2. (i) $\Longrightarrow$ (ii). Suppose $\alpha = \bar{\alpha}$. Let x and $\bar{x}$ be as in the formulation. For the triangle $\triangle abx$ and its comparison triangle $\triangle a_0b_0x_0$ one has $\measuredangle b_0a_0x_0 \ge \alpha = \measuredangle \bar{b}\bar{a}\bar{x}$. Since $|a_0b_0| = |\bar{a}\bar{b}|$ and $|b_0x_0| = |\bar{b}\bar{x}|$, it follows that $|bx| = |b_0x_0| \ge |\bar{b}\bar{x}|$. On the other hand, the curvature condition implies that $|bx| \le |\bar{b}\bar{x}|$. Therefore $|bx| = |\bar{b}\bar{x}|$.

3. (ii) $\implies$ (iii). Suppose that $|bx| = |\bar{b}\bar{x}|$ for some $x \in [a,c] \setminus \{a,c\}$. First we show that the equality $|by| = |\bar{b}\bar{y}|$ holds for all $y \in [a,c]$. Since $|bx| = |\bar{b}\bar{x}|$, the triangles $\bar{b}\bar{x}\bar{a}$ and $\bar{b}\bar{x}\bar{c}$ are comparison triangles for $\triangle bxa$ and $\triangle bxc$. Therefore $\measuredangle bxa \le \measuredangle \bar{b}\bar{x}\bar{a}$ and $\measuredangle bxc \le \measuredangle \bar{b}\bar{x}\bar{c}$. On the other hand, $\measuredangle bxa + \measuredangle bxc \ge \measuredangle bxc = \pi$. Since $\measuredangle \bar{b}\bar{x}\bar{a} + \measuredangle \bar{b}\bar{x}\bar{c} = \pi$, it follows that $\measuredangle bxa = \measuredangle \bar{b}\bar{x}\bar{a}$ and $\measuredangle bxc = \measuredangle \bar{b}\bar{x}\bar{c}$, so the assertion (i) holds for triangles $\triangle bxa$ and $\triangle bxc$. Since we have already proved that (i) implies (ii), we can conclude that $|by| = |\bar{b}\bar{y}|$ for all $y \in [ax] \cup [cx]$.

Now applying the first variation formula (Theorem 4.5.6) to the distance from b along the geodesic $[ac]$, we obtain that $\alpha = \bar{\alpha}$ and $\beta = \bar{\beta}$. The remaining equality $\gamma = \bar{\gamma}$ follows by a combination of the implication (i) $\implies$ (ii) and the already proved part of the implication (ii) $\implies$ (iii).

4. Thus the assertions (i)–(iii) are equivalent. It remains to prove that they imply (iv). The desired totally geodesic surface can be obtained by sweeping it by a family of shortest paths connecting b and $[ac]$.

Let $x, y \in [ac]$. Construct corresponding points $\bar{x}, \bar{x}'$ in $[\bar{a}\bar{c}]$. We assume that $y \in [x,c]$. We have

$$\bar{\beta} \ge \measuredangle \bar{a}\bar{b}\bar{x} + \measuredangle \bar{x}\bar{b}\bar{y} + \measuredangle \bar{y}\bar{b}\bar{c} \ge \measuredangle abx + \measuredangle xby + \measuredangle ybc \ge \beta = \bar{\beta}.$$

Hence all these inequalities turn out to be equalities, in particular, $\measuredangle xby = \measuredangle \bar{x}\bar{b}\bar{y}$. Hence the assertions (i)–(iii) hold for $\triangle xby$ implying that the distances between points of $[bx]$ and points of $[by]$ are the same as between the respective points in $[\bar{b}\bar{x}]$ and $[\bar{b}\bar{y}]$ in the k-plane. Therefore the union of the shortest paths $\{[bx]\}_{x\in[a,c]}$ is a desired "full" triangle isometric to $\triangle \bar{a}\bar{b}\bar{c}$. □

Exercise 9.1.20. Let $\triangle$ be a triangle contained in a normal region of a space of curvature ≤ 0. Let a, b, c denote the lengths of the sides of $\triangle$ and α, β, γ the respective (opposite) angles. Prove the following inequalities:

$$\alpha + \beta + \gamma \le \pi,$$
$$c^2 \ge a^2 + b^2 - 2ab\cos\gamma,$$
$$c \le b\cos\alpha + a\cos\beta.$$

Show that the equality in any of these inequalities implies that the assertions from the above proposition holds.

Hint: Observe that these inequalities turn to equalities for a comparison triangle in $\mathbb{R}^2$; then apply the angle condition.

9.1.5. Reshetnyak's Gluing Theorem. [**Resh**] The following theorem is a simple but very useful tool for verifying upper curvature bounds of "compound" spaces. It also allows us to construct many nontrivial examples of spaces with curvature bounded above.

Theorem 9.1.21. (Reshetnyak's Gluing Theorem.) *Let $\{(X_i, d_i)\}$, $i = 1, 2$, be two (complete locally compact) spaces of curvature $\leq k$. Suppose that there are convex sets $C_i \in X_i$ and an isometry $f\colon C_1 \to C_2$. Attach these spaces together along the isometry f.*

Then the resulting space $(\mathbf{X}, d)$ is a space of curvature $\leq k$.

Remark 9.1.22. Iterating the construction described in Reshetnyak's theorem one can obtain examples glued from more than two (or even countably many) components.

Proof. To simplify notations we will not distinguish spaces X_i and their images in $\mathbf{X}$ under natural projections. In particular, let $C \in \mathbf{X}$ be the common projection of C_i, $i = 1, 2$, to $\mathbf{X}$. Convexity of C implies that restrictions of d to X_i coincide with d_i, $i = 1, 2$. In particular, every point $p \in \mathbf{X}$ has a neighborhood $U \subset \mathbf{X}$ such that $U_i = U \cap X_i$ is a normal ball in X_i. Consider a triangle $\triangle abc$ in U. One may assume that a, $b \in X_1$, $c \in X_2$.

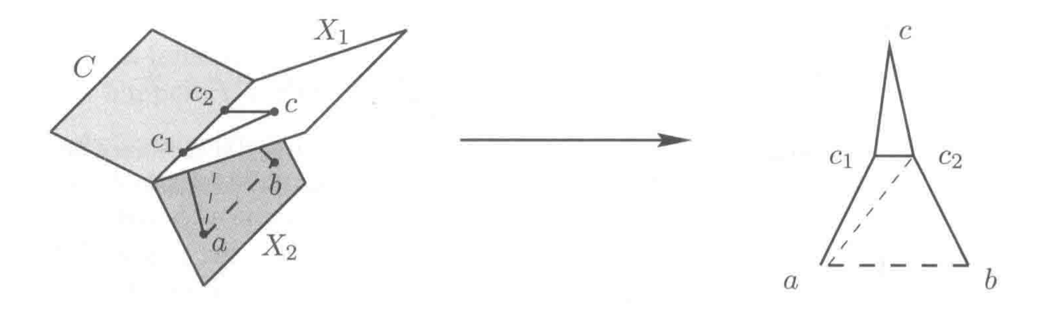


Figure 9.1: Reshetnyak's theorem.

Then there are points c_1, c_2 on the shortest paths $[ac]$, $[bc]$, resp. containing in C. Decompose $\triangle abc$ into three triangles $\triangle ac_1c_2$, $\triangle bac_2$, $\triangle cc_1c_2$. Place their comparison triangles in the plane "in the natural" way (see Figure 9.1). Usual angle comparison arguments show that the angles at c_1 and c_2 in the polyhedron $\bar{a}\bar{c}_1\bar{c}\bar{c}_2\bar{b}$ are concave (i.e., not less than π). Loosely speaking, one can consider this polyhedron like a triangle $\bar{a}\bar{b}\bar{c}$ having two "concave sides", $\bar{a}\bar{c}_1\bar{c}$ and $\bar{b}\bar{c}_2\bar{c}$. "Straightening this triangle" (compare with Alexandov's Lemma 4.3.3) we see that the angles of $\triangle abc$ are not greater than the angles of its comparison triangle. □

9.1.6. Reshetnyak's Majorization Theorem. [**Resh**] There is one more important theorem due to Reshetnyak, the Majorization Theorem. This theorem gives better insight into the nature of spaces of curvature bounded above. We present it here without a proof because we never use it hereafter.

Theorem 9.1.23. (Reshetnyak's Majorization Theorem.) *Let U be a normal ball in a space of curvature not greater k and γ be a closed rectifiable curve in U. If $k > 0$, then we suppose in addition that $L(\gamma) \le \frac{2\pi}{\sqrt{k}}$. Then there exists a closed convex set Ω in the k-plane bounded by a closed curve Γ and a nonexpanding map f of Ω into U such that the restriction $\tilde{f}$ of f on Γ is a length-preserving map onto γ.*

Note that the converse assertion is correct as well: if the statement of the theorem holds for every (sufficiently short) closed curve γ in a length space M, then M is a space of curvature not greater than k. The proof of the latter is simple: it is sufficiently to apply the statement to "triangular" curves.

9.1.7. Extendibility of geodesics. Recall that a geodesic is a curve which is locally a shortest path. We assume that all geodesics are naturally parameterized.

Definition 9.1.24. A geodesic $\gamma\colon [O,a] \to X$ is *extendible* beyond the point $\gamma(a)$ if γ is a restriction of a geodesic $\tilde{\gamma}\colon [0,b] \to X$ with $b > a$.

We will soon see that, loosely speaking, a geodesic in a complete space of nonpositive curvature is always extendible unless its endpoint looks like a boundary point of the space. The following simple statement can help to understand what we mean.

Proposition 9.1.25. *Let X be a complete space of nonpositive curvature. If a geodesic $\gamma\colon [0,a]$ is not extendible beyond the point $p = \gamma(a)$, then the punctured ball $B_\varepsilon(p) \setminus \{p\}$ is contractible for all sufficiently small $\varepsilon > 0$.*

Note that the converse statement is wrong in general: there exists a space having a boundary point such that all geodesics ending at the point are extendible beyond it.

Example 9.1.26. Consider the Euclidean cone over the line. The metric of the cone is locally Euclidean everywhere except the vertex p. Every punctured ball centered at p is contractible, so p is a (unique) boundary point of the cone. However every geodesic ending at p is obviously extendible beyond p.

Exercise 9.1.27. 1. Prove that for a smooth surface in $\mathbb{R}^3$ boundary points are exactly the points that have a contractible punctured neighborhood.

2. Give examples showing that this statement fails for metrics without upper curvature bound.

Proposition 9.1.28. *Let X be a complete metric space and assume that every geodesic is extendible through all of its points. Then every geodesic γ is a restriction of an arc-length parameterized geodesic $\overline{\gamma}\colon \mathbb{R} \to X$.*

In other words, every geodesic can be extended in both sides "to infinity". Note that no curvature restriction is assumed here.

Proof. Actually this fact is a trivial consequence of the Hopf–Rinow Theorem. Let (a, b) be a maximal open interval (possibly infinite) to which γ can be extended. We want to prove that $(a, b) = (-\infty, \infty)$. Suppose for instance that $b < \infty$. Since X is complete, γ can be defined at b by the Hopf–Rinow Theorem. And the assumptions of the proposition imply that γ is extendible beyond b. This contradict the maximality of b. □

Definition 9.1.29. If the conditions of Proposition 9.1.28 hold (i.e., if the space is complete and all geodesics are extendible), we say that X is *geodesically complete.*

Since every geodesic is a shortest path as long as it is contained in a normal ball (see Proposition 9.1.17), we have the following

Corollary 9.1.30. *Let X be a geodesically complete space of curvature bounded above. Then for every point $p \in X$ there is a number $r > 0$ such that all geodesics starting at p can be extended up to the boundary of the ball $B_r(p)$.*

Example 9.1.31. Consider a disc $x^2 + y^2 < 1$ in $\mathbb{R}^2$ and add to it the point $(1, 0)$. This is a "flat" space, and the point $(1, 0)$ is the only point where the conclusion of the corollary fails.

Example 9.1.32. Consider a bouquet of disjoint segments $[0, 1/i]$, $i = 1, 2, \dots$, attached to a point 0. This is a compact space of nonpositive curvature. Ends of the segments are the only points where geodesics are not extendible. And though all geodesics coming to 0 are extendible beyond 0, the point 0 has no round neighborhood mentioned in the corollary above.

9.1.8. Spaces of directions and tangent cones. Space of directions. Let X be a complete locally compact metric space, $p \in X$. Recall (see Subsection 3.6.6) that the space of directions at p is the completion of the metric space Σ'_p whose points are equivalence classes of the shortest paths starting at p and the distance between two classes is the angle between corresponding shortest paths. This space is denoted by Σ_p. In the case of either nonpositively or nonnegatively curved space, angles between shortest paths do exist, so the space of directions is well-defined at every point.

Remark 9.1.33. Even if a space X is geodesically complete, it is possible that Σ'_p is not complete, i.e., some elements of Σ_p are not represented by shortest paths. (This is quite surprising in view of Corollary 9.1.30.)

For example, consider the (open!) half-plane $\{(x, y) : \ y > 0\}$ with the ray $\{(0, y) \colon y \le 0\}$ added. Cut the half-plane along rays $\{(x, y) \colon$

$kx \geq 1,\ y = \frac{x}{k}\}$ for all $k \in \mathbb{Z} \setminus \{0\}$ and glue a copy of the first quadrant $(x, y \geq 0)$ into each cut.

The resulting space is homeomorphic to a half-plane with a ray added, and it is geodesically complete (unlike the bare half-plane or the half-plane with the ray added). The space of directions at $(0,0)$ consists of a single point (corresponding to the vertical ray) and an interval of length π. The endpoints of this interval are not represented by geodesics.

Exercise 9.1.34. Prove the above statements.

However, if a space X is geodesically complete and locally compact, then Σ_p is compact at every point $p \in X$.

Exercise 9.1.35. Prove this.

Hint: Apply Corollary 9.1.30.

Even if X is locally compact and complete (as we usually assume for simplicity), the direction spaces Σ_p are not necessarily compact:

Example 9.1.36. The bouquet of segments considered in 9.1.32 is a compact space of nonpositive curvature. Its space of directions at 0 is not compact—it is a countable discrete space with distances π between different points.

Example 9.1.37. This example is a space of curvature bounded below (not above!), so it rather falls under the subject of Chapter 10. Consider the direct product of countably many round spheres of radii $1/i$ $(i = 1, 2, \dots)$. (Exercise: give a definition of such a product and prove that it is a compact space of nonnegative curvature.)

Its space of directions at each point is not locally compact—it is the product of infinite number of copies of S^1 (exercise: formulate and prove this).

Note that Σ_p is not always connected. The simplest example is the line $\mathbb{R}$: its space of directions at any point consists of two points with the distance π between them. We will need the following

Definition 9.1.38. If every point of a metric space (X, d) has a neighborhood such that restriction of the metric d to this neighborhood is an intrinsic metric, we say that metric d is *locally intrinsic*.

The original (angular) metric of a space of directions Σ_p is not necessarily intrinsic. One of reasons is that angles never exceed π while shortest paths in a space of directions may be much longer than π. For example, consider a cone with the total angle 4π around the origin. Its space of directions at

the origin is a circle of length 4π while all distances (angles) are not greater than π.

So we equip a space of directions not only with its original (angular) metric but also with the intrinsic metric induced by the original one.

As usual, the intrinsic distances may be infinite. This certainly occurs if Σ_p consists of more than one component. Later we will see that in "good" cases this is the only possible reason for the intrinsic metric to take infinite values.

Lemma 9.1.39. *If X is a complete locally compact space of curvature bounded above and $p \in X$, then the angular metric of Σ_p is locally intrinsic.*

More precisely, for every point $a_0 \in \Sigma_p$ and every positive $r < \pi/2$, the restriction of the angular metric to the ball $B_r(a_0)$ is intrinsic.

Proof. By Theorem 2.4.16 it is sufficient to prove that for every two points a, $b \in B_r(a_0)$ and every $\varepsilon > 0$ there is an ε-midpoint. Recall that Σ_p is the completion of its subset Σ'_p consisting of points represented by shortest paths. Without loss of generality, one can assume that points a and b belong to Σ'_p. So they can be represented by shortest paths α and β, $\alpha(0) = \beta(0) = p$. Take an $\varepsilon > 0$ and choose points $x = \alpha(t)$, $y = \beta(t)$ so close to p that $0 \leq \measuredangle\bar{x}\bar{p}\bar{y} - \measuredangle xpy \leq \varepsilon$, where $\measuredangle\bar{x}\bar{p}\bar{y}$ is the angle in a comparison triangle $\triangle\bar{x}\bar{p}\bar{y}$. Let z be a midpoint between x and y. Our assumption guarantees that $d(a,b) < \pi$, where d is the distance in Σ'_p. Hence $z \neq p$ for sufficiently small ε. Place comparison triangles $\triangle\bar{x}\bar{p}\bar{z}$ and $\triangle\bar{y}\bar{p}\bar{z}$ in the k-plane in different half-planes separated by the line $\bar{p}\bar{z}$ and, as usual, compare the resulting quadrilateral with a comparison triangle for the triangle $\triangle xpy$. Since $|\bar{x}\bar{y}| \leq d(x,y)$, we obtain $\measuredangle\bar{x}\bar{p}\bar{z} + \measuredangle\bar{y}\bar{p}\bar{z} \leq \measuredangle xpy + \varepsilon$. With the obvious equality $\measuredangle\bar{x}\bar{p}\bar{z} = \measuredangle\bar{y}\bar{p}\bar{z}$, it gives

$$\measuredangle xpz \leq \measuredangle\bar{x}\bar{p}\bar{z} \leq \frac{1}{2}(\measuredangle xpy + 2\varepsilon), \quad \measuredangle xpz \leq \measuredangle\bar{y}\bar{p}\bar{z} \leq \frac{1}{2}(\measuredangle xpy + 2\varepsilon)$$

if points x and y are sufficiently close to p.

Let c denote the point of Σ'_p represented by a shortest path pz. Now the last two inequalities can be re-written as

$$d(a,c) \leq \frac{1}{2}(d(a,b) + 2\varepsilon), \quad d(b,c) \leq \frac{1}{2}(d(a,b) + 2\varepsilon).$$

By the triangle inequality for angles we have

$$d(a,c) + d(b,c) \geq d(a,b) - \varepsilon.$$

These three inequalities immediately imply that $|d(a,c) - \frac{1}{2}d(a,b)| \leq 2\varepsilon$, $|d(b,c) - \frac{1}{2}d(a,b)| \leq 2\varepsilon$. Since ε is arbitrary, this proves the midpoint property, so the metric of Σ'_p is locally intrinsic. $\square$

Tangent cone. Denote by K_p the cone over the direction space Σ_p (see Subsection 3.6.2 for the definition of the cone over a length space). Every point $w \in K_p$ except its origin o is represented by a pair (ξ, r), where $\xi \in \Sigma_p$ and $r = |ow|$. The cone K_p is called the *tangent cone* at p. It is clear that for a smooth surface in $\mathbb{R}^3$ all tangent cones are planes.

Remark 9.1.40. There is another approach to the notion of tangent cone, namely the Gromov–Hausdorff tangent cone that we discussed in Section 8.2. It is defined as the Gromov–Hausdorff limit of pointed spaces $(\lambda X, p)$ as $\lambda \to \infty$. (Loosely speaking, one "blows up" our metric space X from a fixed point $p \in X$ and calls the limit space the tangent cone.)

In many cases (but not always!) the two definitions are equivalent. One of the obstacles to such equivalence is noncompactness of the space of directions, which brings all difficulties related to Gromov–Hausdorff convergence of not locally compact spaces. Moreover, the limit in the definition of the Gromov–Hausdorff tangent cone may fail to exist even if the space of directions is compact.

Example 9.1.41. Begin with the ray $\mathbb{R}^+ = \{x \in \mathbb{R} : x \geq 0\}$, and attach a segment of length $1/k$ at every point $1/k$, $k \in \mathbb{N}$. The space of directions of the resulting space at 0 is a point, so the corresponding tangent cone is a ray. On the other hand, the Gromov–Hausdorff tangent cone obviously does not exist.

Now let us change this example: attach a segment of length x at *every* point $x \in \mathbb{R}^+$. The resulting space Y is a (not locally compact) space of nonpositive curvature. The direction space of Y at the point 0 still is a point. This time the Gromov–Hausdorff tangent cone at 0 does exist (in the sense of Definition 8.2.1) and is isometric to Y, but it is not a cone over any metric space.

Remark 9.1.42. In general, a small neighborhood of a point in a space of curvature bounded above may have much more complicated structure than the tangent cone. Loosely speaking, the space of directions (and the tangent cone) may carry much less information than a small neighborhood of a point. Contrary to this, in spaces with lower curvature bounds the tangent cone is very similar (at least, homeomorphic) to a small metric ball centered at a given point. This is the reason why the notion of tangent cone (defined either in terms of the space of directions or via the Gromov–Hausdorff limit) is more important for spaces with curvature bounded below.

Let X be a space of curvature bounded above, $p \in X$, $U = B_r(p)$ a normal neighborhood of p. We define two maps, the logarithmic map $\log_p$ and the exponential map $\exp_p$. The map $\log_p$ is a map from U to the tangent cone at p. Let $x \in U$. Connect p to x by a (unique) shortest path $[px]$ and let

$\log_p x = (\xi, r) \in K_p$ where ξ is the direction of $[px]$ and $r = |px|$. It is easy to prove that $\log_p$ is continuous. We would like to define $\exp_p = (\log_p)^{-1}$. However $\log_p$ may be not injective. So we let $\exp_p(x)$ be one (any) point of the set $\log_p^{-1}$. Even in a small neighborhood V of the origin $O \in K_p$ the map $\exp_p$ may not be defined everywhere. However the set where $\exp_p$ is defined is dense in V if V is small enough (prove this!).

If X is a space of nonpositive curvature, then $\log_p$ is a nonexpanding map. And if X is a space of curvature $\le k$, $k > 0$, then $\log_p$ is a Lipschitz map with a constant $1 + \delta$ where δ goes to zero along with the diameter of the neighborhood.

Remark 9.1.43. Now let X be a length space (not necessarily of curvature bounded above), $p \in X$ and K_p well defined. Then we can define maps $\log_p$ and $\exp_p$ almost as above. However now the situation is somewhat different: we have no normal neighborhood of p and cannot guarantee uniqueness of a shortest path $[px]$. So we choose any of such paths. As a result $\log_p$ can be discontinuous (just at points connected with p by two or more shortest paths having different directions at p). Nevertheless we still can define $\exp_p = (\log_p)^{-1}$. And if X is a space of curvature bounded below, $\exp_p$ is continuous and moreover Lipschitz (just like $\log_p$ in the case of curvature bounded above).

Theorem 9.1.44. *If X is a complete locally compact space of curvature bounded above, $p \in X$, then the tangent cone K_p is a length space of nonpositive curvature.*

Proof. The idea of the proof is very simple. One takes a triangle in K_p and considers a sufficiently small homothetic one (with respect to a homothety centered at the origin). The map $\exp_p$ allows us to associate a triangle in X to a small triangle in K_p. These two triangles have almost the same lengths of the respective sides (to the first order), so the curvature comparison condition translates from the triangle in X to the one in K_p.

To make it more clear, let us assume for now that K_p is locally compact and that each point of K_p is represented by a shortest path. Let $\triangle a_1 a_2 a_3$ be a triangle in K_p and a_4 the midpoint of the side $[a_1 a_3]$. Denote by γ_i, $i = 1, 2, 3, 4$, the shortest paths representing points a_i and by a_i^λ the point of the shortest path γ_i such that $|pa_i^\lambda| = \lambda |Oa_i|$. Here O is the origin of K_p and λ is a small positive number. By the very definitions of the angular metric and the cone's metric, we have $\lambda^{-1}|a_i^\lambda a_j^\lambda| \sim |a_j a_j|$ as $\lambda \to 0$ (in the sense that the ratio of the two quantities converges to 1). For a small λ, consider the comparison triangle $\triangle_\lambda$ for $\triangle a_1^\lambda a_2^\lambda a_3^\lambda$ in the k-plane and the dilated triangle $\lambda \triangle_\lambda$. The latter one has sides that converge to those of $\triangle a_1 a_2 a_3$ and lies in the $\lambda^{-1}k$-plane (whose curvature goes to zero as $\lambda \to 0$). Therefore the

curvature comparison conditions for $\triangle a_1^\lambda a_2^\lambda a_3^\lambda$ and a suitable (close to a_4^λ) point of $[a_1^\lambda a_3^\lambda]$ turn in the limit (as $\lambda \to 0$) to the same condition (but with zero curvature bound!) for $\triangle a_1 a_2 a_3$ and a_4. This proves the theorem with our special assumptions.

To get a proof in the general case, use the "generalized" Definition 9.1.2. The same argument works except that a_4 should be an ε-midpoint instead of the real midpoint (which may not exist), and each point a_i should be replaced by a sufficiently close point represented by a shortest path. □

Combining the above theorem with Theorem 4.7.1 about cones over length spaces yields the following

Corollary 9.1.45. *Under the conditions of the theorem, the space of directions Σ_p at every point $p \in X$ is a space of curvature ≤ 1.*

Now we restrict ourselves to geodesically complete spaces (see Definition 9.1.29). We will see now that such spaces enjoy additional nice properties.

Proposition 9.1.46. *Let X be a geodesically complete locally compact length space of curvature bounded above. Then its space of directions at any point is compact.*

Proof. Let $p \in X$ and $B_r(p)$ be a normal neighborhood of p. Consider a metric sphere $S_r(p)$, $0 < r < r'$. The shortest paths connecting p with points of $S_r(p)$ fill in $B_r(p)$ and each geodesic starting at p can be extended to $S_r(p)$ and is a shortest path within $B_r(p)$.

The map associating to each point $x \in S_r(p)$ the direction at p of the unique shortest path $[px]$ is continuous and $S_r(p)$ is compact. So the image of $S_r(p)$ (which is just Σ_p') is compact as well. □

Exercise 9.1.47. Prove that none of the assumptions of the theorem can be omitted.

Theorem 9.1.48. *Let X be a geodesically complete locally compact space of curvature bounded above. Then at every point the Gromov–Hausdorff tangent cone exists and is equal to the tangent cone K_p.*

Proof. Let us assume for simplicity that X is a space of nonpositive curvature. This allows us to write "nonexpanding map" instead of a cumbersome map with a Lipschitz constant $1 + \delta$.

Take two positive numbers, R and ε. By Proposition 7.4.12 we have to find a finite ε-net in $B_R(O) \subset K_p$ which is a uniform (or Lipschitz, here it is the same) limit of ε-nets in balls $B_R^\lambda(p) \subset (\lambda X, p)$ as $\lambda \to \infty$. Choose a closed ball $\overline{B}_r(p)$ in a normal neighborhood of p. By the Hopf–Rinow Theorem 2.5.28 this ball is compact, so there is a finite δ-net in $\overline{B}_r(p)$ for

each $\delta > 0$. Denote by N_0 a finite $\frac{r}{R}\varepsilon$-net in $\overline{B}_r(p)$. Now construct ε-nets N and N_λ in $B_R^\lambda(p)$ and $B_R(0)$, resp. by setting $N = \frac{R}{r}\log_p(N_0)$, where $\frac{R}{r}$ means homothety in K_p.

For each t, $0 < t \leq 1$, consider a map h_t sending every point $x \in \overline{B}_r(p)$ to the point $h_t(x)$ of the shortest path $[xp]$ such that $|h_t(x)p| = t|xp|$. The mapping $h_t : \overline{B}_r(p) \to \overline{B}_{tr}$ is distance nonincreasing and surjective (check this). Now put $N_\lambda = \lambda \circ h_t(N_0)$, where $\lambda\circ$ means re-scaling and $t = \frac{R}{\lambda r}$. For fixed R, λ consider the map

$$\Phi_\lambda = \lambda \circ \log_p \circ \frac{1}{\lambda} : B_R^\lambda(p) \to B_R(0),$$

where $\lambda\colon (X,p) \mapsto (\lambda X, p)$ and $\frac{1}{\lambda}\colon (\lambda X, p) \mapsto (X,p)$ are re-scaling maps. This map does not increase distances and sends N_λ to N. So the only thing we still need is to estimate decrease of distances between the points of N_λ under the map Φ_λ. The set N_0 is finite; thus for every $\nu > 0$, there is a Λ such that for $\lambda > \Lambda$ and $a, b \in N_\lambda$ the angle $\measuredangle\bar{a}\bar{p}\bar{b}$ of a comparison triangle for $\triangle apb$ is not bigger than $\measuredangle apb + \nu$ (since homothety does not change angles).

For $a, b \in N_\lambda$ denote $\tilde{a} = \Phi_\lambda(a)$, $\tilde{b} = \Phi_\lambda(b)$. A comparison triangle $\triangle pab$ and the triangle $\triangle O\tilde{a}\tilde{b}$ have equal lateral sides, $|pa| = |O\tilde{a}|$, $|pb| = |O\tilde{b}|$, and $\measuredangle apb \geq \measuredangle\tilde{a}O\tilde{b} - \nu$, where we can choose $\nu \to 0$ as $\lambda \to \infty$. This proves that $|ab| - |\tilde{a}\tilde{b}| \to 0$ uniformly for all couples $(a,b) \in N_\lambda$ as $\lambda \to \infty$. □

9.2. Hadamard Spaces

For a general (possibly positive) curvature bound k, not much is known about global properties of spaces of curvature $\leq k$. Most known results about such spaces are local. However the assumption that $k \leq 0$ implies deep conclusions about global structure of the space. In particular, if a space is complete and simply connected, it inherits many nice properties of Euclidean spaces and, if $k < 0$, of hyperbolic spaces.

Definition 9.2.1. A complete simply connected space of nonpositive curvature is called a *Hadamard space*.

An important source of examples of Hadamard spaces is the construction of a universal covering (see Subsection 3.4.2 for general discussion of coverings). Let X be a complete space of nonpositive curvature (not necessarily simply connected). Recall (see Remark 9.1.18) that such a space is locally simply connected and therefore has a universal covering; i.e., there exist a simply connected topological space $\tilde{X}$ (the so-called universal covering space) and a covering map $f\colon \tilde{X} \to X$.

The metric of X is (canonically) lifted to $\tilde{X}$ so that the covering map becomes a local isometry. Since being nonnegatively curved is a local property and f is a local isometry, $\tilde{X}$ is nonpositively curved as long as X is. And a covering space of a complete space is complete (cf. Exercise 3.4.8). Thus $\tilde{X}$ is a Hadamard space.

One can apply general properties of Hadamard spaces to the universal covering space $\tilde{X}$ and then derive information about the original space X. This section and Section 9.3 contain a number of statements whose proofs work this way.

9.2.1. Cartan–Hadamard Theorem. We already had examples showing that curvature conditions in spaces of curvature bounded above may fail for large triangles or hinges. Looking at these examples, one observes that they also contain geodesic "bi-angles", i.e., pairs of geodesics (even shortest paths) connecting the same pairs of points. Later we will show that, indeed, if there are no such bi-angles (i.e., if every two points are connected by a unique geodesic), then all curvature conditions hold "in the large" (Theorem 9.2.9).

How does one check that geodesics are unique? In general, this is a hard question, but for nonpositively curved spaces the answer is easy and is given by the following fundamental "generalized Cartan–Hadamard Theorem":

Theorem 9.2.2. *Every two points in a Hadamard space are connected by a unique geodesic. Furthermore, every geodesic segment in a Hadamard space is a shortest path.*

This theorem was proved by E. Cartan for Riemannian manifolds, by M. Gromov for locally compact Hadamard spaces, and by D. Bishop and S. Alexander in the general case (see [**AB**]).

Here we prove it only for locally compact Hadamard spaces. Note that in this case a shortest path connecting any two points definitely exists. So in this case it is sufficient to prove that, for every two points, a geodesic connecting them is unique. For a proof for the general case, see [**AB**] and [**BH**].

Before we begin the proof, it is useful to make some preliminary consideration of geodesics in a space of nonpositive curvature. We assume that geodesics are parameterized with constant speed.

Lemma 9.2.3. *Let α and β be two constant-speed geodesics in a nonpositively curved space parameterized by the same interval and contained in a normal ball. Then the distance $d(\alpha(t), \beta(t))$ is a convex function of t.*

Proof. Denote $d(\alpha(t), \beta(t))$ by $\delta(t)$. Since this function is continuous, it suffices to prove that

$$\delta(\frac{t_1 + t_2}{2}) \le \frac{\delta(t_1) + \delta(t_2)}{2}.$$

We use the following obvious property of nonpositively curved spaces:

For a triangle $\triangle abc$ in such a space and for midpoints b', c' in its sides ab, ac one has $|b'c'| \le |\bar{b}'\bar{c}'|$, where $\bar{b}'$, $\bar{c}'$ are midpoints of the corresponding sides of a comparison triangle $\bar{a}\bar{b}\bar{c}$.

Now let $t = (t_1 + t_2)/2$. Recall that every geodesic segment in a normal ball is a shortest path. Let p be the midpoint between $\alpha(t_1)$ and $\beta(t_2)$. Then

$$\begin{aligned}\delta(t) = |\alpha(t)\beta(t)| &\le |\alpha(t)p| + |p\beta(t)| \\ &\le \frac{1}{2}\Big(|\alpha(t_1)\beta(t_1)| + |\alpha(t_2)\beta(t_2)|\Big) = \frac{1}{2}(\delta(t_1) + \delta(t_2)),\end{aligned}$$

and the lemma follows. □

The next lemma says that, given a geodesic segment γ connecting points p and q, for every point q' near q there exists a unique geodesic connecting p and q' and passing near γ. Note that there may be other geodesics connecting p and q' but passing "far away" from γ; moreover the constructed geodesic may be not a shortest path even if γ is. A reader familiar with Riemannian geometry will notice that this lemma is an counterpart of the fact that geodesics in a nonpositively curved Riemannian manifold have no conjugate points.

Lemma 9.2.4. *Let X be a complete locally compact space of nonpositive curvature, $\gamma\colon [0,1] \to X$ be a constant-speed geodesic with endpoints $p = \gamma(0)$, $q = \gamma(1)$. Let $r > 0$ be so small that the convexity radius of $\gamma([0,1])$ (cf. subsection 9.1.3 for the definition) is greater than $10r$.*

Then for every $q' \in B_r(q)$ there exists a unique constant-speed geodesic $\alpha : [0,1] \to X$ such that $\alpha(0) = p$, $\alpha(1) = q'$, and the uniform distance between γ and α is less than r, i.e., $|\gamma(t)\alpha(t)| < r$ for all t.

Remark 9.2.5. It is easy to see that the geodesic α depends continuously on q' (prove this as an exercise).

The idea of the proof is simple. Consider a class of "broken lines" (that is, curves composed of several shortest paths) with a fixed number of vertices connecting p and q', having short edges and passing in a neighborhood of γ. We will show that a broken line of minimal length in this class is the desired geodesic.

Proof. 1. Divide γ into equal-length intervals by points $p_0 = p$, $p_1 = \gamma(1/N)$, $p_2 = \gamma(2/N)$, ..., $p_N = q$, where N is so large that the length

of each interval is less than r. Consider the class $\mathbf{M}$ of broken lines $a_0a_1a_2\cdots a_N$ such that $a_0 = p$, $a_N = q'$ and $|a_ip_i| \le r$ for all i. By a broken geodesic $a_0a_1a_2\cdots a_N$ we mean a concatenation of shortest paths $[a_ia_{i+1}]$ each constant speed parameterized by the respective interval $[i/N, (i+1)/N]$. Since the space is locally compact, the set of admissible sequences $(a_0, a_1, \dots, a_N)$ is compact. Therefore there is a broken geodesic $\alpha = a_0a_1\dots a_N$ minimizing the length (i.e., the sum $\sum_i |a_ia_{i+1}|$) in the class $\mathbf{M}$.

2. We will show that α is a geodesic; moreover every pair $[a_{i-1}a_i]$ and a_ia_{i+1} of adjacent segments form a shortest paths. Suppose the contrary, i.e., that $|a_{i-1}a_i| + |a_ia_{i+1}| > |a_{i-1}a_{i+1}|$ for some i. Let a_i' be the midpoint of a shortest path $[a_{i-1}a_{i+1}]$. Replacing a_i by a_i' we obtain a broken geodesic which is shorter than α. Thus, in order to obtain a contradiction, it suffices to show that the new broken geodesic belongs to $\mathbf{M}$, i.e., that $|a_i'p_i| \le r$. Apply Lemma 9.2.3 to the interval $[p_{i-1}p_{i+1}]$ of α and the shortest path $[a_{i-1}a_{i+1}]$. Since p_i and a_i' are the midpoints of these intervals, it follows that $|a_i'p_i| \le \frac{1}{2}(|a_{i-1}p_{i-1}| + |a_{i+1}p_{i+1}|) \le \frac{1}{2}(r+r) = r$. Hence the new (shorter) broken geodesic belongs to $\mathbf{M}$. This contradiction shows that α is a geodesic. Moreover, since the distance $|\alpha(t)\gamma(t)|$ is a convex function on every interval $[i/N, (i+1)/N]$ (by Lemma 9.2.3), we have $|\alpha(t)\gamma(t)| \le \max_i |a_ip_i| \le r$.

3. However α is not parameterized with a constant speed (unless all distances $|a_ia_{i+1}|$ are equal). We have to show that α reparameterized with a constant speed remains r-close to γ (with respect to the uniform distance). The argument from the previous step shows that moving a vertex a_i to the midpoint between a_{i-1} and a_{i+1} does not increase the distance between our broken geodesic and γ. It is easy to see that by doing such moves one can obtain a sequence of reparameterizations of α converging to a constant-speed one.

See also a more formal argument in the remark after the proof.

4. It remains to prove the uniqueness of α. Suppose there is another geodesic $\beta : [0,1] \to X$ connecting p and q' in the r-neighborhood of γ. Then the (uniform) distance between α and β is less than $2r$; hence for all $t \in [0,1]$ the points $\alpha(t)$ and $\beta(t)$ belong to some normal ball. By Lemma 9.2.3 it follows that the function $t \mapsto |\alpha(t)\beta(t)|$ is convex and hence has no local maxima in $(0,1)$. On the other hand, this function takes zero values at 0 and 1, so it equals zero everywhere in $[0,1]$. This means that α and β coincide. □

Remark 9.2.6. Here is another argument replacing steps 2 and 3 in the above proof. It is simpler but looks "less natural".

Let us minimize over $\mathbf{M}$ the sum $\sum_i |a_ia_{i+1}|^2$ instead of the bare length. Then the same midpoint construction (replacing a_i by a_i') shows that a

minimizing broken geodesic $a_0a_1 \dots a_N$ is a constant-speed geodesic. Indeed,

$$|a_{i-1}a_i|^2 + |a_ia_{i+1}|^2 \ge 2\left(\frac{|a_{i-1}a_i| + |a_ia_{i+1}|}{2}\right)^2 \ge 2\left(\frac{|a_{i-1}a_{i+1}|}{2}\right)^2 = |a_{i-1}a_i'|^2 + |a_i'a_{i+1}|^2.$$

The equality here is achieved only if $|a_{i-1}a_i| = |a_ia_{i+1}| = \frac{1}{2}|a_{i-1}a_{i+1}|$, i.e., if a_i is a midpoint between a_{i-1} and a_{i+1}. For a minimizing broken geodesic $a_0a_1 \dots a_N$ this holds for all i; hence it is a constant-speed geodesic.

Lemma 9.2.4 can be interpreted as follows. Consider the space X_p of all constant-speed geodesics (parameterized by $[0,1]$) emanating from p. The topology in this space is determined by the uniform distance. Lemma 9.2.4 tells us that for a sufficiently small r the points of the ball $B_r(p)$ are in 1-1 correspondence with the r-neighborhood of γ in the space of geodesics. The correspondence is given by the "exponential" map $\widetilde{\exp}_p$ which sends every geodesic to its endpoint, i.e., $\widetilde{\exp}_p(\alpha) = \alpha(1)$. This map is obviously continuous. Taking into account that X is locally compact, we see that $\widetilde{\exp}_p$ is a local homeomorphism.

Remark 9.2.7. The space X_p may differ from the tangent cone K_P and so $\widetilde{\exp}_p$ may differ from the exponential map $\exp_p$ defined in 9.1.8. They definitely are different if two geodesics emanating from p have the same direction at p and none of them is a part of another.

Now turn to the proof of Theorem 9.2.2.

Proof of Theorem 9.2.2. Let X be a Hadamard space. Fix a $p \in X$ and consider the space X_p of geodesics emanating from p and the map $\widetilde{\exp}_p : X_p \to X$ sending every geodesic to its endpoint (see above). By Lemma 9.2.4, $\widetilde{\exp}_p$ is a local homeomorphism. Lifting locally the metric of X to X_p, we equip X_p with a length metric such that $\widetilde{\exp}_p$ is a local isometry. We will show that $\widetilde{\exp}_p$ is a covering map. The theorem follows easily from this. Indeed, since X is simply connected, the covering map $\widetilde{\exp}_p$ is a homeomorphism; hence for every point $q \in X$ the inverse image $\widetilde{\exp}_p^{-1}(q)$ consists of exactly one geodesic. In other words, there is a unique geodesic connecting p and q.

To prove that $\widetilde{\exp}_p$ is a covering map, we use the criterion provided by Theorem 3.4.18: a local isometry from a complete length space to a "good enough" length space (more precisely, one with locally unique shortest paths) is a covering map. Since X is nonnegatively curved, it is "good enough" in this sense, so it remains to show that X_p (with the length metric lifted from X) is complete.

By Hopf–Rinow Theorem 2.5.28, it is sufficient to prove that, for some $P \in X_p$, any constant-speed geodesics $\Gamma : [0, 1) \to X_p$ emanating from P can be extended to the closed interval $[0, 1]$.

Let P be the constant geodesic (speed 0) resting at p. Then every geodesic $\gamma : [0, 1) \to X$ with $\gamma(0) = p$ has a unique lift $\Gamma : [0, 1) \to X_p$ with $\Gamma(0) = P$. (Being a lift means that $\gamma = \widetilde{\exp}_p \circ \Gamma$.) This lift Γ is defined in a natural way: for every $t \in [0, 1)$, the point $\Gamma(t)$ in the space of geodesics is nothing but the (reparameterized) restriction $\gamma|_{[0,t]}$. The uniqueness of a lift follows from the fact that $\widetilde{\exp}_p$ is a local homeomorphism.

It is easy to extend such a lift Γ to $[0, 1]$. Extend γ to $[0, 1]$ (this is possible since X is complete) and take the lift of the extended curve. In other words, define $\Gamma(1) = \gamma|_{[0,1]}$. Since every geodesic in X_p is a lift of some geodesic in X (namely, of its own image under $\widetilde{\exp}_p$), we have proved the desired sufficient condition for X_p to be complete. The theorem then follows as explained above. □

A geodesic loop is a geodesic $\gamma : [a, b] \to X$ such that $\gamma(a) = \gamma(b)$. Closed (or periodic) geodesics are particular case of geodesic loops.

Corollary 9.2.8. *Let X be a complete locally compact space of nonpositive curvature and $p \in X$. Then every element of the fundamental group $\pi_1(X, p)$ contains exactly one geodesic loop.*

Proof. Consider a universal covering map $f : \tilde{X} \to X$ and lift the metric of X to $\tilde{X}$. Then $\tilde{X}$ is a Hadamard space. Fix a point $x \in f^{-1}(p)$. For every closed curve γ in X with endpoints at p there is a unique lift $\tilde{\gamma}$ in $\tilde{X}$ starting at X; furthermore $\tilde{\gamma}$ is a geodesic if and only if γ is. Two curves represent the same element of the fundamental group if and only if the endpoints of their lifts coincide. So the set of curves representing a given element of $\pi_1(X, p)$ corresponds to the set of curves in $\tilde{X}$ connecting x and some fixed point $y \in f^{-1}(p)$. The latter contains exactly one geodesic by Cartan–Hadamard Theorem; hence $\pi_1(X, p)$ contains exactly one geodesic loop. □

9.2.2. Globalization.

Theorem 9.2.9. *Let X be a Hadamard space of curvature $\leq k$. Then the curvature conditions hold for all triangles in X; i.e., X is a space of curvature $\leq k$ "in the large".*

Remark 9.2.10. In fact, the condition that X is a Hadamard space can be replaced by the conclusion of the Hopf–Rinow Theorem: every two points are connected by a unique shortest path and the shortest paths depend continuously on the endpoints.

Proof of the theorem. We will prove that the angle condition holds for all triangles. (This implies that all other curvature conditions hold globally, because the "global" versions of the definitions of bounded curvature are equivalent just like the local ones.)

1. The key observation is the following: if $\triangle abc$ is a triangle in X, d is a point in its side $[ac]$, and the angle condition holds for the triangles $\triangle abd$ and $\triangle cbd$, then it holds for $\triangle abc$.

Indeed, place comparison triangles $\triangle\bar{a}\bar{b}\bar{d}$ and $\triangle\bar{c}\bar{b}\bar{d}$ for $\triangle abd$ and $\triangle cbd$ in the k-plane in different half-planes with respect to their common side $\bar{b}\bar{d}$. From the angle condition for $\triangle abd$ and $\triangle cbd$ we have $\measuredangle bda \le \measuredangle\bar{b}\bar{d}\bar{a}$ and $\measuredangle bdc \le \measuredangle\bar{b}\bar{d}\bar{c}$; hence $\measuredangle\bar{b}\bar{d}\bar{a} + \measuredangle\bar{b}\bar{d}\bar{c} \ge \measuredangle bda + \measuredangle bdc \ge \pi$. Therefore the angle at $\bar{d}$ of the quadrilateral $\bar{a}\bar{b}\bar{c}\bar{d}$ is not less than π. Using Alexandrov's Lemma 4.3.3 we obtain that $\measuredangle bad \le \measuredangle\bar{b}\bar{a}\bar{d} \le \bar{b}_1\bar{a}_1\bar{c}_1$, where $\triangle\bar{b}_1\bar{a}_1\bar{c}_1$ is a comparison triangle for $\triangle bac$. Similarly $\measuredangle bcd \le \measuredangle\bar{b}_1\bar{c}_1\bar{a}_1$. Finally,

$$\measuredangle abc \le \measuredangle abd + \measuredangle cbd \le \measuredangle\bar{a}\bar{b}\bar{d} + \measuredangle\bar{c}\bar{b}\bar{d} = \measuredangle\bar{a}\bar{b}\bar{c} \le \measuredangle\bar{a}_1\bar{b}_1\bar{c}_1$$

(the last inequality follows from the fact that $|\bar{a}\bar{c}| \le |\bar{a}\bar{d}| + |\bar{c}\bar{d}| = |\bar{a}_1\bar{c}_1|$).

2. We say that a triangle $\triangle abc$ is *slim* if its edges $[ab]$ and $[ac]$ are very close to each other; more precisely, the (uniform) distance between those sides is less than $r/10$ where r is the convexity radius of $[ab] \cup [ac]$ (so every r-ball centered in a point of $[ab] \cup [ac]$ is a normal region).

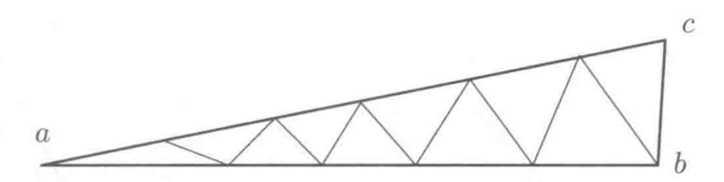


Figure 9.2: Cutting a slim triangle.

Let us "cut" such a slim triangle $\triangle abc$ into small triangles as shown in Figure 9.2. Each small triangle is contained in a normal ball and therefore satisfies the angle condition. Then the angle condition for $\triangle abc$ follows by induction on the number of small triangles from the fact proved in the first step.

3. Thus the angle condition is satisfied for all slim triangles. Now consider an arbitrary triangle $\triangle abc$. Since shortest paths in X are unique and depend continuously on their endpoints (recall that we suppose M to be locally compact), we have a continuous family of shortest paths connecting a to the points of the side $[b, c]$. Then one can partition $[bc]$ into small intervals $[b_i b_{i+1}]$, $i = 0, 1, \dots, N-1$, $b_0 = b$, $b_N = c$, so that each triangle $\triangle ab_ib_{i+1}$ is slim. These slim triangles satisfy the angle condition. Then the

result of the first step implies by induction on N that $\triangle abc$ satisfies the angle condition. □

Remark 9.2.11. Conversely, if a complete space has nonpositive curvature "in the large", then it is simply connected and hence is a Hadamard space.

Indeed, suppose that such a space X is not simply connected. Fix a point $a \in X$ and find a shortest nontrivial geodesic loop γ with endpoint at a. Then divide γ by points b and c into three intervals with equal lengths. These intervals are shortest path; hence we have a triangle $\triangle abc$ whose sides form a single geodesic. This triangle violates the angle condition.

9.2.3. Convex functions.

Definition 9.2.12. Let X be a length space. A function $f : X \to \mathbb{R}$ is said to be *convex* if its restriction to any constant-speed geodesic γ (that is, the function $t \mapsto f(\gamma(t))$) is convex.

If the function f is continuous and the metric of X is strictly intrinsic, convexity of f is equivalent to the following: for any points $x, y, z \in X$ such that z is a midpoint between x and y, one has $f(z) \le \frac{1}{2}(f(x)+f(y))$. Indeed, this midpoint criterion is equivalent to requiring that f is convex along any shortest path, and the latter implies convexity along any geodesic (because being a convex function is a local property).

In Hadamard spaces, many natural functions are convex. First, the globalization theorem allows us to "globalize" Lemma 9.2.3:

Proposition 9.2.13. *Let X be a Hadamard space. Then for any two constant-speed geodesics α and β the function $\delta(t) = d(\alpha(t), \beta(t))$ is convex.*

Corollary 9.2.14. *If X is a Hadamard space, then for every $p \in X$ the function $x \mapsto |px|$ is convex.*

Proof. Let β be a constant geodesic staying at p, and apply Proposition 9.2.13. □

Corollary 9.2.15. *Let X be a Hadamard space and $Y \subset X$ be a convex set. Then the distance from Y (that is, the function $x \mapsto \operatorname{dist}(x, Y)$) is convex.*

Proof. We assume that X is locally compact. Let $x, y \in X$, z be a midpoint between x and y, x', y' be points nearest to x and y in the closure of Y, and z' be a midpoint between x' and y'. Since Y is convex, z' belongs to the closure of Y. Then $\operatorname{dist}(z, Y) \le |zz'| \le \frac{1}{2}(|xx'| + |yy'|)$. Here the second inequality follows from Proposition 9.2.13. □

Corollary 9.2.16. *Let X be a Hadamard space and $i \colon X \to X$ an isometry of X. Then the "displacement function" $\delta(x) = d(x, i(x))$ is convex.*

Proof. The i-image of a geodesc γ is a geodesic $i \circ \gamma$. And Proposition 9.2.13 says that the restriction of δ to γ, i.e., the function $d(\gamma(t), i \circ \gamma(t))$ is convex. □

The notion of λ-convex functions was introduced in Section 4.4, Example 4.4.4. Recall that λ-convex functions are in a sense "no less convex" than a quadratic function $t \mapsto \lambda t^2$. Roughly speaking, these are functions whose second derivatives (in some generalized sense) are bounded below by 2λ. More precisely:

Definition 9.2.17. Let X be a length space and $\lambda > 0$. A function $f : X \to \mathbb{R}$ is λ-*convex* if for any unit-speed geodesic γ in X the function $t \mapsto f(\gamma(t)) - \lambda t^2$ is convex.

A function is said to be *strongly convex* if it is λ-convex for some $\lambda > 0$.

Here is a reformulation of the definition in terms of midpoints.

Proposition 9.2.18. *A continuous function $f : X \to \mathbb{R}$ is λ-convex if and only if, for any $x, y \in X$ and z a midpoint between x and y,*

$$f(z) \le \frac{f(x) + f(y)}{2} - \frac{\lambda}{4}|xy|^2.$$

Proof. The proof is a straightforward computation. Let γ be a unit-speed geodesic, and denote $g(t) = f(\gamma(t)) - \lambda t^2$. Convexity of g is equivalent to requiring that $g(t_1) + g(t_2) \ge 2g(\frac{1}{2}(t_1 + t_2))$ for all t_1, t_2. If $t = \frac{1}{2}(t_1 + t_2)$, $x = \gamma(t_1)$, $y = \gamma(t_2)$, $z = \gamma(t)$, this inequality can be rewritten as

$$f(x) - \lambda t_1^2 + f(y) - \lambda t_2^2 \ge 2f(z) - 2\lambda t^2,$$

or, equivalently,

$$f(x) + f(y) - 2f(z) \ge \lambda(t_1^2 + t_2^2 - 2(\tfrac{t_1+t_2}{2})^2) = 2\lambda(\tfrac{t_1-t_2}{2})^2 = \frac{\lambda}{2}|xy|^2$$

which is just the inequality given in the formulation. □

The next theorem characterizes Hadamard spaces in terms of convexity of distance functions. In fact, this theorem is nothing but a (globalized) reformulation of the definition of nonpositive curvature (compare with Section 4.4).

We use the following notation: for a metric space X and a point $p \in X$, d_p denotes the distance function of p, that is, the function defined by $d_p(x) = |px|$.

Theorem 9.2.19. *For any point p in a Hadamard space, the function d_p^2 is 1-convex.*

Conversely, if X is a complete space with strictly intrinsic metric and d_p^2 is 1-convex for any $p \in X$, then X is a Hadamard space.

Proof. In the Euclidean plane, the function d_p^2 has the form $t \mapsto t^2 + \text{const}$ along a unit-speed straight line. Since in a Hadamard space the distance function is "more convex" than in the plane (this is the distance condition, our first definition of nonpositive curvature), d_p^2 is "more convex" than t^2.

Here is the same argument filled in with formulas. By Proposition 9.2.18, it is sufficient to verify that

$$|pz|^2 \le \frac{|px|^2 + |py|^2}{2} - \frac{|xy|^2}{4}$$

if z is a midpoint between x and y. In the Euclidean plane this inequality turns to equality. (This is the formula for the length of a triangle's median in terms of sides.) By the distance condition, the median $[pz]$ of $\triangle pxy$ is not longer than the respective median of a comparison triangle, and the desired inequality follows.

Conversely, if d_p^2 is convex for any point p, the above inequality for the median holds for all triangles. This means that X has nonpositive curvature "in the large". This implies that X is a Hadamard space, see Remark 9.2.11 after the proof of Globalization theorem. $\square$

One of the most useful properties of strongly convex functions is that they always attain minima.

Proposition 9.2.20. *Let X be a complete space with a strictly intrinsic metric, and $f : X \to \mathbb{R}$ a continuous strongly convex function bounded from below. Then f has a unique minimum point.*

Proof. The uniqueness part is easy. If $f(x) = f(y) = \min f$ and $x \ne y$, let z be a midpoint between x and y, and apply Proposition 9.2.18. This yields $f(z) < \frac{1}{2}(f(x) + f(y)) = \min f$, contradiction.

Now let us prove that a point of minimum exists. Let $m = \inf f$ and $\{x_i\}_{i=1}^{\infty}$ be a minimizing sequence for f, i.e., $f(x_i) \to m$ as $i \to \infty$. We are going to show that $\{x_i\}$ is a Cauchy sequence.

Fix an $\varepsilon > 0$ and let i_0 be so large that $f(x_i) < m + \varepsilon$ for all $i > i_0$. Then for all $i, j > i_0$ and $z \in X$ we have

$$f(z) \ge m > \frac{f(x_i) + f(x_j)}{2} - \varepsilon.$$

Taking a midpoint between x_i and x_j for z and subtracting the inequality from Proposition 9.2.18, we obtain that $\lambda |x_i x_j|^2/4 < \varepsilon$. Thus $|x_i x_j| < \sqrt{4\varepsilon/\lambda}$ for all $i, j > i_0$. Since λ is fixed and ε is arbitrarily small, it follows that $\{x_i\}$ is a Cauchy sequence.

Now define $x = \lim x_i$. Then $f(x) = \lim f(x_i) = m$, so x is the desired minimum point. $\square$

Many geometric constructions in Hadamard spaces are based on minimum points of convex functions. Below are some examples.

Definition 9.2.21. Let $\{p_i\}_{i=1}^n$ be a collection of points in a metric space and $\{m_i\}_{i=1}^n$ a collection of positive numbers. A *barycenter* (or a *center of mass*) of points $\{p_i\}$ with masses $\{m_i\}$ is the minimum point of the function $\sum m_i d_{p_i}^2$ where $d_{p_i}(x) = |p_i x|$.

In a Hadamard space, the functions $d_{p_i}^2$ are strongly convex (Theorem 9.2.19), and so is the function $\sum m_i d_{p_i}^2$. Therefore a barycenter exists and is unique.

Exercise 9.2.22. Prove that the barycenter depends continuously on the points $\{p_i\}$ and masses $\{m_i\}$, and is a Lipschitz function of $\{p_i\}$ if the masses m_i are fixed.

Definition 9.2.23. Let X be a metric space and $Y \subset X$ a bounded set. A (closed) ball of minimal radius among the balls containing Y is called a *circumscribed ball* of Y, its center a *circumcenter* of Y, and its radius the *circumradius*.

Proposition 9.2.24. *In a Hadamard space, every bounded set has a unique circumscribed ball.*

Proof. A circumcenter of Y is a minimum point of a function R_Y defined by $R_Y(x) = \sup\{|xy| : y \in Y\}$. Let us minimize the function R_Y^2 instead. This function is the supremum of functions d_y^2 over $y \in Y$. A supremum of 1-convex functions is obviously 1-convex. So R_Y^2 is 1-convex and therefore a minimum point of R_Y^2 exists and unique. □

Exercise 9.2.25. Consider the space $\mathfrak{M}(X)$ of all bounded subsets of X equipped with Hausdorff distance. Prove that, if X is a Hadamard space, then the circumcenter is continuous as a function on $\mathfrak{M}(X)$.

9.2.4. Parallel rays and lines. Recall that a *line* in a length space X is a (unit-speed) geodesic $\gamma : \mathbb{R} \to X$ such that any closed subinterval of γ is a shortest path. A *ray* is a geodesic $\gamma : \mathbb{R}_+ \to X$ possessing the same property. In Hadamard spaces, every geodesic segment is a shortest path, so lines and rays are just complete geodesics of infinite length.

As in Euclidean and hyperbolic spaces, one can introduce a notion of parallel rays and lines in a Hadamard space.

Definition 9.2.26. Two rays or two lines are said to be *parallel* if the uniform distance between them is finite. Parallel rays are called asymptotic as well.

It is obvious that being parallel is an equivalence relation.

Exercise 9.2.27. Prove that two rays or lines are parallel if and only if the Hausdorff distance between them is finite.

In the Euclidean plane two rays or lines are parallel if and only if they are parallel in ordinary ("scholar") sense. It is easy to see that if two lines in the Lobachevsky plane are parallel, then they coincide (but parallel rays starting at points a, b in the Lobachevsky plane do exist for every pair of points a, b). We will see soon that in the general case the situation is very similar.

Proposition 9.2.28. *Let X be a Hadamard space and $p \in X$. Then for every ray γ in X there exists a unique ray starting at p and parallel to γ.*

Proof. We prove this fact only for locally compact spaces. Let γ_t denote the shortest path connecting p to $\gamma(t)$. Since the distance between geodesics is a convex function (Proposition 9.2.13), this shortest path is contained in the (closed) r-neighborhood of γ where $r = |p\gamma(0)|$. Since the space is locally compact, one can choose a sequence $\{t_i\}$, $t_i \to \infty$, such that $\{\gamma_{t_i}\}$ converges to some ray $\alpha : \mathbb{R}_+ \to X$. This ray stays within the r-neighborhood of γ and hence is parallel to γ.

To prove the uniqueness, suppose that there are two rays α and β starting at p and parallel to γ. Since being parallel is an equivalence relation, the rays α and β are parallel, so the function $t \mapsto |\alpha(t)\beta(t)|$ is bounded. On the other hand, this function is convex and equals zero at $t = 0$; therefore it is zero for all t. This means that $\alpha = \beta$. $\square$

With the notion of parallel rays, one defines the *ideal boundary* of a Hadamard space similarly to the ideal boundary of the Lobachevsky plane. Namely, the points of the ideal boundary are equivalence classes of parallel rays. The ideal boundary of a Hadamard space X is denoted by $X(\infty)$. It is common to say that a ray "connects" its initial point with the point of the ideal boundary that it represents. In this language, the above proposition tells us that every point of X and every point of $X(\infty)$ are connected by a unique ray.

The ideal boundary and various structures on it provide a useful language for describing asymptotic properties of a Hadamard space. In this book we do not get further into this important subject (see [**BGS**], [**E**], [**BH**] for some further material and references).

Now we turn to parallel lines. It turns out that parallel lines do not exist unless a space contains some flat subsets; in particular, there are no parallel lines in Hadamard spaces of strictly negative curvature.

Theorem 9.2.29. *Let X be a Hadamard space, and γ_1 and γ_2 be two parallel lines. Then either γ_1 and γ_2 coincide, or they span a convex flat strip.*

Therefore if X has curvature $\le k$ for a $k < 0$, then there are no parallel lines in X except coinciding ones.

By a flat strip spanned by γ_1 and γ_2 we mean a convex subset isometric to a strip between two parallel lines in $\mathbb{R}^2$ via an isometry which maps the boundary lines to γ_1 and γ_2.

For the proof we need the following simple lemma.

Lemma 9.2.30. *Let $Q = abcd$ be a quadrangle in Hadamard space and the sum of the angles of Q no less than 2π. Then this sum of angles equals 2π, and Q is a boundary of a flat convex region isometric to a planar quadrangle.*

Proof. Draw the "diagonal" $[ac]$ in the quadrangle Q. The angle condition for nonpositive curvature implies that the sums of angles of $\triangle abc$ and $\triangle adc$ are not greater than π. On the other hand, the triangle inequality for angles implies that the sum of all six angles in $\triangle abc$ and $\triangle adc$ is not less than the sum of angles of Q and hence is no less than 2π. Therefore all mentioned inequalities turn to equalities; in particular, the sum of angles of Q equals 2π, and in both triangles $\triangle abc$ and $\triangle acd$ the sum of angles equals π. Therefore (by Proposition 9.1.19) each of these triangles bounds a totally geodesic flat surface (a "solid" triangle). Denote these surfaces by T_1 and T_2, respectively. We are going to show that $T_1 \cup T_2$ is a desired quadrangle.

By a similar argument, the triangles $\triangle abd$ and $\triangle cbd$ bound flat solid triangles T_3 and T_4. To finish the proof, it is sufficient to show that $T_1 \cup T_2 = T_3 \cup T_4$, or, equivalently, that the shortest path $[ac]$ lies in the surface $T_1 \cup T_2$. Observe that both surfaces $T_1 \cup T_2$ and $T_3 \cup T_4$ are arcwise-isometric to a quadrangle $\bar{a}\bar{b}\bar{c}\bar{d}$ whose sides and angles equal the respective sides and angles of Q. Then the length of $[ac]$ equals the diagonal $\bar{a}\bar{c}$, and this diagonal corresponds to a curve of the same length in $T_1 \cup T_2$. Thus there is a curve in $T_1 \cup T_2$ whose length equals $|ac|$. This means that this curve is a shortest path and, since shortest paths in a Hadamard space are unique, coincides with $[ac]$. □

Proof of the theorem. Let α and β be two parallel lines. By Proposition 9.2.13, the function $t \mapsto |\alpha(t)\beta(t)|$ is convex on $\mathbb{R}$. Since this function is bounded, it is a constant. Note that this remains true if one shifts the parameterization of α or β by a change of variable $t \mapsto t + \text{const}$. Let us choose a parameterizations so that the value of $|\alpha(t)\beta(t)|$ is the minimum possible. To do this, let $\beta(0)$ be a point nearest to $\alpha(0)$ in the line β. Then for every $t \in \mathbb{R}$ we have

$$|\alpha(t)\beta(t)| = |\alpha(0)\beta(0)| = \min_{c\in\mathbb{R}} |\alpha(0)\beta(c)| = \min_{c\in\mathbb{R}} |\alpha(t)\beta(t+c)| = \min_{c\in\mathbb{R}} |\alpha(t-c)\beta(t)|,$$

so the point $\beta(t)$ is nearest to $\alpha(t)$ in β, and the point $\alpha(t)$ is nearest to $\beta(t)$ in α. Then by the first variation formula (Theorem 4.5.6) all four angles that a shortest path $[\alpha(t)\beta(t)]$ forms with half-lines of α and β are no less than $\pi/2$. Hence the quadrangle $\alpha(0)\beta(0)\beta(t)\alpha(t)$ satisfies the assumption of Lemma 9.2.30, so its angles equal $\pi/2$ and it bounds a flat surface. The union of such rectangles over all $t \in \mathbb{R}$ is the desired flat strip. □

Now we can prove a splitting theorem for Hadamard spaces.

Theorem 9.2.31. *Let X be a Hadamard space, γ a line in X, and $Y \subset X$ be a union of lines parallel to γ. Then Y is isometric to a product $Z \times \mathbb{R}$ for some metric space Z.*

In particular, if every point of X belongs to a line parallel to γ, then X is isometric to a product $Z \times \mathbb{R}$ where Z is a Hadamard space.

Proof. If α and β are two parallel lines, $x \in \alpha$, and $y \in \beta$ is a nearest point to x in β, we call y a *projection* of x to β.

Let us choose a parameterization of every line α parallel to γ so that $\alpha(0)$ is a projection of $\gamma(0)$ to α. We first prove that, for every two lines α and β parallel to γ, the point $\beta(0)$ is the projection of $\alpha(0)$ to β. Suppose this is not the case. Let $p = \alpha(0)$, and let $q = \beta(t_0)$ be a projection of p to β. We may assume that $t_0 > 0$; otherwise change the orientation of the lines. By Theorem 9.2.29 the lines α and β bound a flat strip of width $|pq|$; hence $|\beta(t)p| = \sqrt{|\beta(t)q|^2 + |pq|^2}$ for all $t \in \mathbb{R}$. It follows that $|\beta(t)p| - |\beta(t)q| \to 0$ as $t \to \pm\infty$ (because $|\beta(t)q| \to \infty$ and $|pq|$ is fixed). Then

$$|\beta(t)p| - t = |\beta(t)p| - |\beta(t)\beta(0)| = |\beta(t)p| - |\beta(t)q| + t_0 \xrightarrow[t\to+\infty]{} t_0.$$

Therefore $|\beta(t)p| > t + t_0/2$ for all large enough t. We are going arrive at a contradiction by showing that there is a path from $\beta(t)$ to p (passing through a point of γ) whose length is less than $t + t_0/2$.

By Theorem 9.2.29 the lines γ and α bound some flat strip S_1, and lines γ and β bound some flat strip S_2. Consider the union $S = S_1 \cup S_2$ of these strips. It is arcwise-isometric to a planar strip whose width is the sum of widths of S_1 and S_2 (more precisely, S is the image of this planar strip under an arcwise isometry). Furthermore, the segment between points corresponding to $p = \alpha(0)$ and $\beta(0)$ in the planar strip is orthogonal to its boundary lines. Reasoning as above, we conclude that

$$t - |\beta(t)p|_s = |\beta(t)\beta(0)|_s - |\beta(t)p|_s \xrightarrow[t\to+\infty]{} 0$$

where $|\cdot\cdot|_s$ is the intrinsic metric of S. Therefore $|\beta(t)p|_s < t_0/2$ for all large enough t. This contradicts the inequality $|\beta(t)p| > t_0/2$ that we had above, because the intrinsic metric of S is not less than the metric of X.

This contradiction proves that $\beta(0)$ is the projection of $\alpha(0)$ to β; in other words, the relation of being a projection is transitive. Now it is easy to finish the proof. Let Z be the set of all projections of $\gamma(0)$ to the lines of which Y is composed. There is a natural map from $Z \times \mathbb{R}$ to Y which maps a pair (z, t) to the point $\gamma_z(t)$ where γ_z is the line parallel to γ passing through z (recall that $z = \gamma_z(0)$ by our choice of parameterizations). Theorem 9.2.29, along with the fact that points of Z are projections of one another to their respective lines, implies that this map is an isometry. □

9.3. Fundamental Group of a Nonpositively Curved Space

In this section we prove some classical results about fundamental groups of nonpositively curved spaces. If X is a compact space of strictly negative curvature (that is, of curvature $\le k$ for a $k < 0$), the results will actually strengthen the fact that these groups are hyperbolic. For instance, let us mention one corollary in advance: for X as above every nontrivial abelian subgroup of $\pi_1(X)$ is isomorphic to $\mathbb{Z}$.

Let us recall the basic facts from Subsection 3.4.2 about fundamental groups and universal coverings. Let X be a length space and $f\colon \tilde{X} \to X$ be the universal covering map. Then the fundamental group $\pi_1(X)$ acts on $\tilde{X}$ by isometries; more precisely, $\pi_1(X)$ is isomorphic to the group of deck transformations, that is, of maps from $\tilde{X}$ to itself commuting with f.

In other words, the isometry group $Iso(\tilde{X})$ contains a subgroup isomorphic to $\pi_1(X)$; moreover this subgroup acts freely (i.e., every nontrivial element of this subgroup is a map without fixed points) and totally discontinuously.

If X is a complete nonpositively curved space, then $\tilde{X}$ is a Hadamard space. Thus, instead of fundamental groups of such spaces X, one can study groups acting by isometries on a Hadamard space. In this case the action is assumed to be free and totally discontinuous.

Theorem 9.3.1. *Let X be a Hadamard space. Then every finite subgroup Γ of $Iso(X)$ has a fixed point; i.e., there exists an $x \in X$ such that $gx = x$ for all $g \in \Gamma$.*

Proof. For every point $p \in X$ its orbit $\Gamma p = \{\gamma(p) : \gamma \in \Gamma\}$ is finite. By Proposition 9.2.24, there is a unique circumcenter p of the orbit. For every $\gamma \in \Gamma$ the map $x \to \gamma(x)$ sends any orbit of Γ onto itself. Therefore $\gamma(p) = p$ for every $\gamma \in \Gamma$. □

Since deck transformations have no fixed points, the theorem yields the following

Corollary 9.3.2. *Let X be a complete space of nonpositive curvature. Then the fundamental group $\pi_1(X)$ does not contain nontrivial finite subgroups.*

In other words, every element of $\pi_1(X)$ except the identity generates an infinite subgroup.

Theorem 9.3.3 (Pressmann). *Let X be a compact space of curvature $\le k$ where $k < 0$. Then every two commuting elements of $\pi_1(X,p)$ belong to a cyclic subgroup.*

Proof. Suppose that $a_0, b_0 \in \pi_1(X,p)$ and $a_0b_0 = b_0a_0$ If one of elements a_0, b_0 is trivial or $a_0 = b_0$, then there is nothing to prove. So suppose that this is not the case. Consider the free homotopy class of the element a_0 considered as maps $S^1 \to X$. There is a shortest representative in this class. (The existence of a shortest loop is proved similarly to the existence of shortest paths in a compact space; see Subsection 2.5.2.) Such a path α is a closed (periodic) geodesic (check this). Pick a point q in α and consider α as a loop with the vertex q. There is an isomorphism of $\pi_1(X,p)$ onto $\pi_1(X,q)$ sending a_0 to $[\alpha] = a$. Denote by b the image of b_0 under this isomorphism. Obviously $ab = ba$.

Let $f : \tilde{X} \to X$ be a universal covering map. A lift of α is the complete geodesic $\tilde{\alpha}$. Let us identify $\pi_1(X,q)$ with the deck transformations group Γ. The element a of Γ sends the geodesic $\tilde{\alpha}$ into itself and acts in $\tilde{\alpha}$ as a group $\mathbb{Z}$ of shifts. (This element cannot change orientation of $\tilde{\alpha}$; otherwise there would be a fixed point in $\tilde{\alpha}$.) The isometry b sends $\tilde{\alpha}$ to a geodesic $\tilde{\beta}$.

First of all we want to show that $\tilde{\beta} = \tilde{\alpha}$. Suppose it is not the case. Take a point x of the geodesic $\tilde{\alpha}$, and consider the quadrangle Q with the vertices x, $x_1 = ax$, $x_2 = bx$ and $x_3 = abx = bax$.

Since a, b are isometries, we have equalities

$$\measuredangle x_1xx_2 + \measuredangle xx_2x_3 = \pi, \ \measuredangle xx_1x_3 + \measuredangle x_1x_3x_2 = \pi.$$

Now it follows from Lemma 9.2.30 that Q is isometric to a planar quadrangle. This contradicts the assumption that X has strictly negative curvature. Therefore $\tilde{\beta} = \tilde{\alpha}$.

Now consider restrictions of the actions of a, b to $\tilde{\alpha}$. These restrictions may be identified with numbers $\bar{a}, \bar{b} \in \mathbb{R}^1$. The group generated by these numbers, that is, the lattice $\{m\bar{a} + n\bar{b}\}_{m,n\in\mathbb{Z}}$, cannot have accumulation points, i.e., is discrete.

It is a well-known fact of the number theory that in this case there is a number $\bar{c}$ such that $\bar{c} = p\bar{a} + q\bar{b}$ and $\bar{a} = r\bar{c}$, $\bar{b} = s\bar{c}$ where p, q, r, s are

integers, $rp + sq = 1$. (To prove this, observe that the set of sums $pa + qb$ assumes its minimum which is just the desired c.)

Let p, q be as above, and consider the element $c = a^p b^q \in \Gamma$. We show that $a = c^r$, $b = c^s$. Indeed, ac^{-r} keeps $\tilde{\alpha}$ fixed and so ac^{-r} is the identity transformation. The proof of the equality $b = c^s$ is the same. □

Combining this theorem with the previous corollary, we obtain

Corollary 9.3.4. *Let X be a compact space of curvature $\le k$ where $k < 0$. Then every nontrivial abelian subgroup of $\pi_1(X)$ is isomorphic to $\mathbb{Z}$.*

Generalizations. The above theorem fails if one drops the assumption about strictly negative curvature. Tori and direct products $T^n \times M$, where M is a Hadamard space, are good counter-examples. In fact, every such counter-example contains a subspace isometric to a flat torus. To get an idea why this happens, recall the considerations of parallel lines in Subsection 9.2.4, in particular Theorem 9.2.29. Flat tori in a nonpositively curved space correspond to abelian subgroups in the fundamental group. More precisely, the following theorem holds.

Theorem 9.3.5. *Let X be a compact space of nonpositive curvature. If $\pi_1(X)$ contains an abelian subgroup G of rank $k > 1$, then X contains a convex subset isometric to a k-dimensional flat torus T.*

Moreover, this torus T can be chosen so that the inclusion map $T \hookrightarrow X$ induces an injective homomorphism of the fundamental groups and G is the image of this homomorphism.

This theorem is a particular case of more general results on the action of properly discontinuous isometry groups on a Hadamard space (see [**BH**] for details and further references).

9.3.1. Final remarks. This chapter is only an introduction to the upper curvature bounded world. The reader can find more results in several recent books and papers. First of all, there is [**BH**] where, besides many other things, the reader can find a detailed exposition of the fundamental group properties (compare with 9.3) and of the geometry of the boundary at infinity—the areas which are hardly touched in our textbook (see also [**BGS**], [**Gro3**]). Also polyhedral complexes of curvature bounded above (spaces are built of simplexes in a space form of curvature k) are considered there in detail. Since 2-dimensional polyhedral spaces are already a very interesting case, we mention here also the paper [**BB**], where one can find appropriate references and some generalization and unsolved problems.

Local structure and some asymptotic properties of Alexandrov spaces of curvature bounded above were studied recently in [**Kl**]. This study is

based on the notion of *geometric dimension* $\dim_G$. It is defined inductively as follows: $\dim_G(X) = 0$ if X is a discrete space and $\dim_G(X) = 1 + \sup_{p \in X} \dim_G(\Sigma_p)$ in other cases. B. Kleiner proved that geometric dimension is equal to topological dimension $\dim_{Top}$ in the following sense: $\dim_G(X) = \sup \dim_{Top}(K)$ where supremum is taken over all compact sets $K \subset X$. If $\dim_G(X) < \infty$, then X has nice properties; for instance, for every $\varepsilon > 0$ there exists a $(1+\varepsilon)$-bi-Lipschitz map of an open (nonempty) set $U \subset \mathbb{R}^n \to X$ where $n = \dim_G(X)$.

For Riemannian manifolds, some aspects of Hadamard spaces are more developed than in the general case (see [**BGS**], [**E**]).

9.4. Example: Semi-dispersing Billiards

Everybody has seen a billiard table with several balls moving and colliding in it. Here we demonstrate how the theory of Alexandrov spaces can be applied to a certain class of problems in the theory of semi-dispersing billiard systems.

Apparently, one of the motivations to study semi-dispersing billiard systems comes from gas models in statistical physics. For instance, the hard ball model is a system of round balls moving freely and colliding elastically in empty space or in a box. Physical considerations naturally lead to several mathematical problems regarding the dynamics of such systems. For instance, one of the central problems of this kind asks for upper bounds on the number of collisions that could occur in a billiard system (in a given time). A series of basic problems of this type in their "physical" version go back to Boltzmann, while their rigorous mathematical study was initiated by Ya. Sinai.

As a (leading) example, consider a system of N round balls moving freely and colliding elastically in empty space $\mathbb{R}^3$ (or in a box). Every ball moves along a straight line with constant speed until two balls collide, and then the new velocities of the two balls are determined by the (conservation) laws of classical mechanics. (Explicit formulas for the velocities after a collision in terms of the masses, radii, and the initial velocities of the balls are usually discussed even in high-school physics classes.) To simplify our considerations, we consider only situations where at most two balls collide simultaneously.

A position of a collection of N balls can be represented by a point in $\mathbb{R}^{3N}$. Namely if $a_i \in \mathbb{R}^3$ is the center of the i-th ball and (x_i, y_i, z_i) are its Cartesian coordinates, the corresponding point in $\mathbb{R}^{3n}$ is

$$(a_1, a_2, \dots, a_N) = (x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_n, y_n, z_n).$$

Not every point in $\mathbb{R}^{3N}$ represents a valid configuration of balls. We have to exclude positions where some of the balls overlap. The ith and jth ball intersect if $|a_i - a_j| < r_i + r_j$ where r_i and r_j are the radii of the balls. This inequality defines a cylinder $C_{ij} \subset \mathbb{R}^{3n}$. The complement $\mathbb{R}^{3N} \setminus \bigcup_{i \neq j} C_{ij}$ is the *configuration space* of our system. Its points correspond to valid positions of the system of balls.

It is natural to consider the configuration space $\mathbb{R}^{3N}$ with a Euclidean structure given by the kinetic energy of the system:

$$K((v_1, v_2, \ldots, a_N), (v_1, v_2, \ldots, v_N)) = \sum_{i=1}^{N} m_i \langle a_i, a_i \rangle ,$$

where $\langle , \rangle$ is the usual scalar product for each of the "three-dimensional" coordinates, and m_i is the mass of the ith ball. The evolution of the system of balls traces a path in the configuration space. It is easy to verify that the point representing the configuration of balls moves straight and at a constant speed until it hits one of the cylinders C_{ij} (this event corresponds to a collision in the system of balls), and then it continues following the standard law of billiard collision: the angle of reflection is equal to the angle of incidence. In other words, when a trajectory hits a cylinder C_{ij}, the projection of the velocity vector onto (the plane tangent to) the cylinder does not change, and the normal component of the velocity vector changes its sign. Of course, both the projection and the normal component are considered with respect to K.

Thus the dynamics of a system of several balls is (equivalent to) a particular case of a semi-dispersing billiard system, which is defined as follows.

Definition 9.4.1. Let M be a complete nonpositively curved Riemannian manifold with nonzero injectivity radius, and $\{B_i\}_{i \in I}$ be a finite (or at least locally-finite) collection of convex subsets of M with smooth boundaries $W_i = \partial B_i$ (which are smooth convex hypersurfaces). The (closure of) the complement $M \setminus \bigcup B_i$ is called a *semi-dispersing billiard table.* The hypersurfaces W_i (as well as the convex sets B_i) will be referred to as *walls* of the table.

A *billiard trajectory* is a unit-speed piecewise smooth curve ("broken line") in the closure of $M \setminus \bigcup B_i$ with break point at walls (i.e., in $\bigcup W_i$), such that the smooth intervals are Riemannian geodesics, and at every break point the left and right velocities have equal projections onto the tangent plane to the wall. In other words, it is a trajectory of a point that moves along a geodesic until it hits one of the walls W_i, and then it gets reflected according to the standard law of ellastic collision. For simplicity, we exclude

the trajectories that ever experience a collision with more than one wall simultaneously.

An informal idea that a semi-dispersing billiard system is somewhat similar to a geodesic flow on a negatively curved manifold has been around for quite a while. Perhaps it was first explicitly mentioned by V. Arnold. In the early sixties V. Arnold "speculated" that "such systems can be considered as the limit case of geodesic flows on negatively curved manifolds (the curvature being concentrated on the collisions hypersurface)". Indeed, it is nowadays well known that a large portion of the results in the smooth theory of (semi-)hyperbolic systems can be generalized (with appropriate modifications) to (semi-)dispersing billiards. In spite of this, the construction suggested by Arnold has never been used until recently. It also caused several serious objections; in particular, A. Katok pointed out that such approximations by geodesic flows on manifolds necessarily produce geodesics that "bend around" collision hypersurfaces and therefore have no analogs in the billiard system.

To study the billiard flow for a fixed time and in a small neighborhood of a fixed point, one can use *doubling* by taking two copies of M, removing the interiors of the walls and then gluing the copies along the boundaries of the walls. One can approximate the singular metric resulting from this procedure by smooth metrics (analogously one substitutes hard collisions by a very steep repelling potential). The geodesic flows of the resulting metrics will naturally converge to the billiard flow on a fixed time interval in a small neighborhood of each point. Even though this construction does not seem very useful, it already can deliver certain information. For instance, the Liouville theorem (invariance of the Liouville measure) for billiard flows follows immediately from the Liouville theorem for geodesic flows.

Let us illustrate Arnold's suggestion by a simple example of the billiard table in the complement of a disc in a two-torus (or the Euclidean plane). Starting with two copies of the torus with (open) discs removed, and gluing them along the boundary circles of the discs, one obtains a Riemannian manifold (a surface of genus 2) with a metric singularity along the gluing circle. This manifold is flat everywhere except at this circle. One can think of this circle as carrying singular negative curvature, and indeed this is a nonpositively curved length space. Smoothing this metric by changing it in an (arbitrarily small) collar around the circle of gluing, one can obtain a smooth nonpositively curved Riemannian metric, which is flat everywhere except in this collar. To every segment of a billiard trajectory, one can (canonically) assign a geodesic in this metric. Collisions with the disc would correspond to intersections with the circle of gluing, where the geodesic leaves one copy of the torus and goes to the other one.

Exercise 9.4.2. Repeat the same construction in dimension three: consider two copies of $\mathbb{R}^3$ with a unit ball removed, and glue them along the boundary spheres. Show that this is a space of curvature bounded above by 1, and it is *not* a nonpositively curved space!

Hint: In this space, the gluing locus (which is a sphere) is a totally geodesic subspace.

Show that the presence of positive curvature in this example persists under smoothenings that do not change the metric outside some neighborhood of the gluing locus.

This exercise shows that there are serious problems with using this construction in higher dimensions. Moreover, even in dimension two many geodesics do not correspond to billiard trajectories. They can be described as coming from "fake" trajectories hitting the disc at zero angle, following an arc of its boundary circle (possibly even making several rounds around it) and then leaving it along a tangent line. Dynamically, such geodesics carry "the main portion of entropy" and they cannot be disregarded. On the other hand, it is difficult to tell actual trajectories from the fake ones when analyzing the geodesic flow on this surface.

Main construction. The purpose of this section is to give an informal and elementary account of how Arnold's idea can nonetheless be formalized (and the difficulties mentioned above can be partially avoided) by using Alexandrov spaces. Several important open problems in the area were recently solved by this method.

Our discussion is concentrated around the idea of gluing several copies of M together and then developing billiard trajectories into this new space. This idea is very old and its simplest versions arise even in elementary high-school mathematical puzzles. For instance, if the billiard table is a square, one can consider a tiling of Euclidean plane by such squares, and billiard trajectories turn into straight lines.

We are going to represent billiard trajectories by geodesics in a nonpositively curved space. To demonstrate the power of this method, we will show how metric geometry allows to solve the problem of estimating the number of collisions. Let us mention that, unfortunately, a construction that would allow us to represent all billiard trajectories as geodesics in one compact space is unknown in dimensions higher than three. Attempts to do this lead to a striking open question: Is it possible to glue finitely many copies of a regular 4-simplex to obtain a (boundary-less) nonpositive pseudo-manifold?

We introduce a construction that represents trajectories from a certain combinatorial class, where by a combinatorial class of (a segment of) a billiard trajectory we mean a sequence of walls that it hits.

Fix such a sequence of walls $K = \{W_{n_i}, i = 1, 2, \ldots, N\}$. Consider a sequence $\{M_i, i = 0, 1, \ldots, N\}$ of isometric copies of M. For each i, glue M_i and M_{i+1} along B_{n_i}. Since each B_{n_i} is a convex set, the resulting space M_K has the same upper curvature bound as M due to Reshetnyak's Theorem 9.1.21.

There is an obvious projection $M_K \to M$, and M can be isometrically embedded into M_K by identifying it with one of M_i's (regarded as subsets of M_K). Thus every curve in M can be lifted to M_K in many ways. A billiard trajectory whose combinatorial class is K admits a canonical lifting to M_K: we lift its segment till the first collision to $M_0 \subset M_K$, the next segment between collisions to $M_1 \subset M_K$ and so on. It is called the development of a trajectory. It is easy to see that a development of a trajectory is a geodesic in M_K.

Note that, in addition to several copies of the billiard table, M_K contains other redundant parts formed by identified copies of B_i's. For example, if we study a billiard in a curved triangle with concave walls, B_i's are not the boundary curves. Instead, we choose as B_i's some convex ovals bounded by extensions of these walls. (One may think of a billiard in the compact component of the complement to three discs.) In this case, these additional parts look like "fins" attached to our space. In case of the billiard in the complement of a disc in a two-torus (see discussion above), the difference is that we do *not* remove the disc when we glue together two copies of the torus. Now a geodesic cannot follow an arc of the disc boundary, as the latter can be shortened by pushing inside the disc. Still, there are "fake" geodesics, which go through the disc. However, there are fewer of them than before and it is easier to separate them.

It might seem more natural to glue along the boundaries of W_{n_i} rather than along the whole B_{n_i}. For instance, one would do so thinking of this gluing as "reflecting in a mirror" or by analogy with the usual development of a polygonal billiard. However, gluing along the boundaries will not give us a nonpositively curved space in any dimension higher than 2 (see Exercise 9.4.2).

One may wonder how the interiors of B_i's may play any role here, as they are "behind the walls" and billiard trajectories never get there. For instance, instead of convex walls in a manifold without boundary, one could begin with a manifold with several "convex" boundary components (with a nonnegative definite second fundamental forms with respect to the inner normal). Even for one boundary component, there are examples where it is impossible to "fill in" the boundary by a nonpositively curved manifold. Moreover, our main dynamical result does fail for an example of this sort. Thus, it is

indeed important that the walls are not only locally convex surfaces, and we essentially use the fact that they are filled by convex bodies.

Estimating the number of collisions. Now we are turning to the main application of our construction, estimating the number of collisions in a hard ball gas model. For the hard ball system, one asks whether the number of collisions that may occur in this system can be estimated from above by a bound depending only on the number of balls and their masses. If we consider the balls moving in unbounded Euclidean space, we count the total number of collisions in infinite time. For a system of balls in a box, we mean the number of collisions in unit time (for a fixed value of kinetic energy). We will consider a general case of a semi-dispersing billiard system. In this case, it is clear that an additional assumption is needed. Indeed, already for a two-dimensional billiard table bounded by several concave walls, a trajectory may experience an arbitrarily large number of collisions (in time one) in a neighborhood of a vertex if two boundary curves are tangent to each other. Thus, we want to assume that (in a certain sense) our billiard table is nondegenerate.

For simplicity, let us introduce the following *nondegeneracy assumption*, which rules out various degenerations of the arrangements of hyperplanes tangent to walls (it can be essentially weakened for noncompact billiard tables):

- there exists a number C such that, for all sufficiently small ε, if a point is ε-close to all sets from some sub-collection I of the B_i's, then it is $C\varepsilon$-close to the intersection of the B_i's from this sub-collection.

Exercise 9.4.3. Verify this condition for a system of several balls in $\mathbb{R}^3$.

Hint: This condition means that for every configuration such that, for any pair of indices from a certain set of pairs I, the corresponding balls are close, there is a nearby configuration for which the balls from each of the pairs intersect. This can be shown by using the following procedure: Choose a ball and move all the balls simultaneously and with equal velocities along the segments connecting their centers with the center of the chosen ball.

For a system of balls in a jar with concave walls the nondegeneracy condition is satisfied except for some special sets of radii, when it is possible to "squeeze the balls tightly between the walls." Actually, it is known that in those situations the system may have arbitrarily many collisions locally.

Now the main local result reads as follows:

Theorem 9.4.4. *If a semi-dispersing billiard table satisfies the nondegeneracy assumption, then there exists a finite number P such that every point p*

in the billiard table possesses a neighborhood $U(p)$ such that every trajectory segment contained in $U(p)$ experiences no more than P collisions.

Passing to estimating the global number of collisions (for infinite time) we want to stay away from situations such as a particle infinitely bouncing between two disjoint walls:

Theorem 9.4.5. *If a semi-dispersing billiard table satisfies the nondegeneracy assumption, M is simply connected and the intersection $\bigcap B_i$ of B_i's is nonempty, then there exists a finite number P such that every trajectory experiences no more than P collisions.*

Of course, for a hard ball gas model, the cylinders C_{ij} have nonempty intersection: for instance, the origin $(0, 0, \ldots, 0) \in \mathbb{R}^{3N}$ belongs to all C_{ij} (note that this point obviously does not correspond to any physically realizable configuration of balls). One can check that the maximal number of collisions that may occur in a system of N hard elastic balls (of arbitrary masses and radii) never exceeds

$$\left(400N^2 \frac{m_{max}}{m_{min}}\right)^{2N^4},$$

where m_{max} and m_{min} are, correspondingly, the maximal and the minimal masses in the system.

Exercise 9.4.6. A *trivial question*: How many collisions can occur in a system of two balls of equal masses in $\mathbb{R}^3$? A *tricky question*: How many collisions can occur in a system of three balls of equal masses in $\mathbb{R}^3$?

Answer: four—and it is not that easy even to give an example...

Sketch of Proof for Theorem 9.4.4 for two walls: To outline the idea of the proofs of uniform estimates on the number of collisions we restrict ourselves to the simplest nontrivial case where the result was unknown: two walls W_1 and W_2 bounding two convex sets B_1 and B_2. Thus we avoid inessential combinatorial complications and cumbersome indices. The reader can understand where the difficulty lies by thinking of a particle shot "almost parallel to the intersection line of W_1 and W_2" and bouncing between them experiencing "almost tangential collisions". The problem is to show that the number of such collisions can be estimated by a bound independent of how closely the particle follows the intersection line.

Let us assume that M is simply connected; otherwise, one can pass to its universal cover. Consider a billiard trajectory γ connecting two points x and y and choose any point $z \in B_1 \cap B_2$. Denote by $K = \{W_1, W_2, W_1, W_2, \ldots\}$ the combinatorial class of γ, and consider the development $\tilde{\gamma}$ of γ in M_K. This is a geodesic between two points x' and y'. By the Cartan–Hadamard

Theorem 9.2.2 every geodesic in a complete simply connected nonpositively curved space is the shortest path between its endpoints. Note that z canonically lifts to M_K since all copies of z in different copies of M got identified. Denoting this lift by z', we see that $|zx| = |z'x'|$ and $|zy| = |z'y'|$. Thus we conclude that the lengths of γ between x and y is less that $|xz|+|zy|$ for all $z \in B_1 \cap B_2$. In other words, any path in M connecting x and y and visiting the intersection $B_1 \cap B_2$ is longer than the segment of γ between x and y.

The following argument is the core of the proof. It shows that if a trajectory made too many collisions then it can be modified into a shorter curve with the same endpoints and passing through the intersection $B_1 \cap B_2$. This contradicts the previous assertion and thus gives a bound on the number of collisions.

Assume that γ is contained in a neighborhood $U(p)$ and it collided with W_1 at points $a_1, a_2, \dots, a_N$ alternating with collisions with W_2 at $b_1, b_2, \dots, b_N$. Let z_i be the point in $B_1 \cap B_2$ closest to b_i and h_i be the distance from b_i to the shortest geodesic $[a_i a_{i+1}]$. By the nondegeneracy assumption, $|z_i b_i| \le C \cdot \mathrm{dist}(b_i, B_1) \le C h_i$. Thus the distance H_i from z_i to the shortest geodesic $[a_i a_{i+1}]$ is at most $(C+1)h_i$.

Plugging this inequality between the heights of the triangles $\triangle a_i b_i a_{i+1}$ and $\triangle a_i z_i a_{i+1}$ into a routine argument which develops these triangles on both the Euclidean plane and k-plane, one concludes that $D_i \le C_1 \cdot d_i$, where $d_i = |a_i b_i| + |b_i a_{a+1}| - |a_i a_{i+1}|$, $D_i = |a_i z_i| + |z_i a_{a+1}| - |a_i a_{i+1}|$. Here k is the infimum of the sectional curvature in $U(p)$, and a constant C_1 can be chosen depending on C alone provided that $U(p)$ is sufficiently small.

Let d_j be the smallest of d_i's. Let us modify the trajectory γ into a curve with the same endpoints: substitute its pieces $a_i b_i a_{i+1}$ by the shortest segments $[a_i a_{i+1}]$ for all i's excluding $i = j$. This new curve is shorter than γ by at least $(N-1)d_j$. Let us make a final modification by replacing the piece $a_j b_j a_{j+1}$ by $a_j z_j a_{j+1}$. It makes the path longer by D_j, which is at most $C_1 d_j$. Hence, $N \le C_1 + 1$ because otherwise we would have a curve with the same endpoints as γ, passing through $z_j \in B_1 \cap B_2$ and shorter than γ. This proves the local bound on the number of collisions. □

Let us pass to the proof of global estimates, where geometry works in its full power. Again, to stay away from combinatorial complications, we restrict ourselves to the case of two walls W_1 and W_2.

Sketch of Proof for Theorem 9.4.5 for two walls. Consider a trajectory γ making N collisions with the (only possible) sequence of walls $K = \{1, 2, 1, \dots, 2, 1\}$. Reasoning by contradiction, assume that $N > 3P+1$, where P is the local bound on the number of collisions (whose existence is

guaranteed by Theorem 9.4.4). Again consider the space M_K, but now we will "close it up" by gluing $M_0 \in M_K$ and $M_N \in M_K$ along the copies of B_1. Denote the resulting space by $\tilde{M}$. We cannot use Reshetnyak's Theorem 9.1.21 directly to conclude that $\tilde{M}$ is a nonpositively curved space any more, since we identify points in the same space and we do not glue two spaces along a *convex* set.

We recall that a space has nonpositive curvature if and only if every point possesses a neighborhood such that, for every triangle contained in the neighborhood, its angles are no bigger than the corresponding angles of the comparison triangle in the Euclidean plane. However, using the correspondence between geodesics and billiard trajectories, one can conclude (reasoning exactly as in the proof of the *local* estimates on the number of collisions) that each side of a small triangle cannot intersect interiors of more than P copies of the billiard table. Since $N > 3P + 1$, for every small triangle for which we want to verify the angle comparison property, we can undo one of the gluings without tearing the sides of the triangle. This ungluing may only increase triangle's angles, and now we find ourselves in a nonpositively curved space (which is actually just M_K), and thus we get the desired comparison for the angles of the triangle.

To conclude the proof, it remains to notice that the development of γ in $\tilde{M}$ is a geodesic connecting two points in the same copy of B_1. This is a contradiction since every geodesic in a simply connected nonpositively curved space is the only shortest path between its endpoints; on the other hand, there is a shortest path between the same points going inside this copy of B_1. □

Chapter 10

Spaces of Curvature Bounded Below

This is an introduction to the theory of length spaces with lower curvature bounds. We consider only complete length spaces (in fact, local completeness is sufficient in all local statements) of curvature $\geq k$ for some $k \in \mathbb{R}$. Throughout the chapter, the term "Alexandrov space" is reserved for (connected) spaces from this class.

In the case of positive curvature bound k it is convenient to exclude some exceptional one-dimensional spaces. Namely, we do not count the line $\mathbb{R}$, the half-line $\mathbb{R}_+$, segments of length greater that $\pi/\sqrt{k}$, and circles of length greater than $2\pi/\sqrt{k}$ as Alexandrov spaces of curvature $\geq k$. As we will see in Section 10.4, excluding these spaces is equivalent to requiring that the diameter of a space is not greater than $\pi/\sqrt{k}$. Thus the term "Alexandrov space of curvature $\geq k$" in this chapter means a (connected) complete length space of curvature $\geq k$ which is not isometric to one of the above listed exceptions if $k > 0$. A developed theory for such spaces was created during 1980–90s. Here we present only some basic results from the foundations of this theory, and a few "global" results that generalize classical theorems of Riemannian geometry. More advanced exposition of results and techniques can be found in [**BGP**], [**Pl**]; for other references and applications see [**GP**].

There are two main points making the theory very different from the case of curvature bounded above. First, Toponogov's Globalization Theorem (Theorem 10.3.1) says that in any Alexandrov space of curvature $\geq k$ (for any k, and without any topological assumptions like simple connectedness),

the triangle comparison conditions hold "in the large". This yields a number of results about global geometry of Alexandrov spaces (recall from Chapter 9 that one can say much more about a Hadamard space than about a general space of curvature bounded above). Furthermore, Toponogov's Theorem implies that the class of Alexandrov spaces of curvature $\geq k$ is closed with respect to Gromov–Hausdorff convergence and this allows us to study this class "as a whole" (recall the discussion in the beginning of Chapter 7).

Second, (finite-dimensional) Alexandrov spaces have nice *local* structure. By dimension we mean Hausdorff dimension (cf. Subsection 1.7.4) but in fact all known notions of dimension are equivalent for an Alexandrov space (in particular, Hausdorff dimension equals topological dimension). Some of the local properties of finite-dimensional Alexandrov spaces are worth mentioning right now. Hausdorff dimension of such a space is always an integer; such a space is a manifold (in fact, almost a Riemannian one) everywhere except a tiny set of singular points; the space of directions of an n-dimensional Alexandrov space is an $(n-1)$-dimensional Alexandrov space of curvature ≥ 1 and in fact is isometric to S^{n-1} at almost every point.

Though these formulations are nice and clear, known proofs are long and technical. We do not get into these details until the end of the chapter (Sections 10.8 and 10.9). Some knowledge of the local structure is, however, required in earlier sections; in such cases we refer to the appropriate results from the last sections.

10.1. One More Definition

In Chapter 4 we gave several equivalent definitions of a space of nonnegative curvature. Replacing $\mathbb{R}^2$ by the k-plane, one obtains definitions of a space of curvature $\geq k$ where k is an arbitrary real number (cf. Section 4.6). Again, these definitions are equivalent and the proof of this equivalence is almost the same as in the case $k = 0$.

Here we add one more equivalent formulation that has no counterpart for spaces of curvature bounded above. Its main feature is logical simplicity; it has the form "for every four points (in a small neighborhood) the distances between them satisfy certain inequalities". Recall that the definitions we had so far involve collections of points satisfying additional restrictions (e.g., one is a midpoint for two others), or even angles.

The new equivalent formulation is given in Proposition 10.1.1. For convenience we introduce the following notation for comparison angles, generalizing it from Chapter 3.

Notation. Let a, b and c be three different points in a length space. We denote by $\widetilde{\angle}abc$ (comparison angle) the angle at $\bar{b}$ of a comparison triangle

$\bar{a}\bar{b}\bar{c}$ in the k-plane (assuming k is fixed). In case of ambiguity we add k as a subscript: $\tilde{\measuredangle}_k abc$.

Note that $\tilde{\measuredangle}_k abc$ is a (continuous) function of the three distances $|ab|$, $|ac|$ and $|bc|$. It is well-defined if $|ab| + |ac| + |bc| < 2R_k$ where R_k is the diameter of the k-plane.

Proposition 10.1.1. *A locally compact length space X is a space of curvature not less k if and only if every point $x \in X$ has a neighborhood U such that for any collection of four different points $a, b, c, d \in U$ the following condition is satisfied:*

$$\tilde{\measuredangle}_k bac + \tilde{\measuredangle}_k cad + \tilde{\measuredangle}_k dab \leq 2\pi. \tag{10.1}$$

We call the inequality (10.1) the *quadruple condition* (a quadruple is a collection of four different points) for a quadruple $(a; b, c, d)$. Note that the condition is symmetric in $\{b, c, d\}$ but a is in a special position. Proposition 10.1.1 allows us to use the quadruple condition as yet another definition of a space of curvature $\geq k$. Note that this new definition does not rely on shortest paths, so it can be used unmodified for not strictly intrinsic spaces.

Proof of Proposition 10.1.1. First suppose that (10.1) holds for any quadruple. To verify the triangle condition (from Definition 4.1.9) for a triangle $\triangle abc$ and $d \in [ac]$, apply the quadruple condition to $(d; a, b, c)$. Since $\tilde{\measuredangle} adc = \pi$, it follows that $\tilde{\measuredangle} bdc + \tilde{\measuredangle} bda \leq \pi$. Then Lemma 4.3.3 implies the desired inequality $|db| \geq |\bar{d}\bar{b}|$.

Now let X be a space of curvature $\geq k$. Consider a quadruple $(a;\ b, c, d)$ and a point a' in a shortest path $[ab]$. Then

$$\begin{aligned} \tilde{\measuredangle} ba'd + \tilde{\measuredangle} da'c + \tilde{\measuredangle} ca'b &\leq \measuredangle bpa'd + \measuredangle dpa'c + \measuredangle cpa'b \\ &\leq (\measuredangle ba'd + \measuredangle da'a) + (\measuredangle aa'c + \measuredangle cpa'b) = 2\pi. \end{aligned}$$

We used the angle condition (Definition 4.1.15), the triangle inequality for angles, and the fact that the sum of adjacent angles equals π (Lemma 4.3.7). Now let a' converge to a; then (10.1) follows by continuity of comparison angles. □

We leave as exercises several simple facts about spaces of curvature $\geq k$.

Exercise 10.1.2. Prove that geodesics in a space of curvature $\geq k$ do not branch. Namely, if two geodesics have a common interval, then they are subintervals of one geodesic.

Exercise 10.1.3. Show that the quadruple condition implies the following fact: for any three geodesics emanating from one point, the sum of three angles between these geodesics is not greater than 2π.

Exercise 10.1.4. Prove that, if two geodesics in a space of curvature $\geq k$ start at one point and form a zero angle at that point, then one of them is a subinterval of the other.

Hint: Use the angle condition to obtain a local statement; then refer to Exercise 10.1.2.

Exercise 10.1.5. Prove that, if shortest paths in a space of curvature $\geq k$ have two points in common, then these points are their endpoints.

10.2. Constructions and Examples

There are many constructions of Alexandrov spaces of curvature bounded below. A number of simple ones were presented in Chapter 4. Here we add two new examples: quotients by isometry groups and convex surfaces. In the second example, the proof of the fact that the space has nonnegative curvature is not trivial and relies on results that we will prove later.

10.2.1. Products and cones.

Example 10.2.1 (products). Let X and Y be Alexandrov spaces of curvature $\geq k$, $k \leq 0$. Then the direct product $X \times Y$ is a space of curvature $\geq k$. The proof is straightforward.

Note that a product of spaces of curvature $\geq k$ is *not* a space of curvature $\geq k$ if $k > 0$ unless one of the multiplied spaces is a single point. See comment just after 9.1.5.

Indeed, consider two arbitrary shortest paths in spaces X and Y. These shortest paths (as subspaces of X and Y) are isometric to intervals of $\mathbb{R}$; hence their product in $X \times X$ is a convex set isometric to a region in $\mathbb{R}^2$. Therefore $X \times Y$ cannot have strictly positive curvature.

Example 10.2.2 (cones). Euclidean and spherical cones were already considered in Subsections 3.6.2, 3.6.3. Here we give a uniform description for these constructions and define the similar notion of a *hyperbolic cone*. In the context of spaces with lower curvature bounds, it does not make sense to consider cones over spaces of diameter greater than π.

Let $k \in \mathbb{R}$ and X be a metric space with $\operatorname{diam}(X) \leq \pi$. The *k-cone* over X, denoted by $\operatorname{Con}_k(X)$, consists of the origin (or apex) o and pairs (x, r) where $x \in X$ and $r > 0$ (in addition, $r \leq \pi/\sqrt{k}$ if $k > 0$). The distance from (x, r) to the origin equals r, and then the distance between $a_1 = (x_1, r_1)$ and $a_2 = (x_2, r_2)$ is defined so that

$$\widetilde{\measuredangle}_k a_1 o a_2 = |x_1 x_2|.$$

In other words, $|a_1 a_2|$ equals the side $|\bar{a}_1 \bar{a}_2|$ of a triangle $\bar{a}_1 \bar{o} \bar{a}_2$ in the k-plane such that $|\bar{o}\bar{a}_i| = r_i$ for $i = 1, 2$, and $\measuredangle \bar{a}_1 \bar{o} \bar{a}_2 = |x_1 x_2|$.

If $k > 0$, all pairs $(x, \pi/\sqrt{k})$ should be identified (because the distance between them is zero). These pairs represent a point $o' \in \mathrm{Con}_k(X)$ which could be taken as the origin instead of o (the map $(x, r) \mapsto (x, \pi/\sqrt{k} - r)$ is an isometry of $\mathrm{Con}_k(X)$). In this case the k-cone is also called the k-spherical cone or the k-suspension. The origins o and o' are referred to as its *poles*. The standard spherical suspension defined in Subsection 3.6.3 corresponds to $k = 1$.

If $k < 0$, k-cones are also called $(-k)$-hyperbolic cones. A *hyperbolic cone* is a k-cone for $k = -1$. If $k = 0$, k-cones are ordinary ("Euclidean") cones.

By the very definition, $\mathrm{Con}_k(S^1)$ is isometric to the k-plane. Similarly, $\mathrm{Con}_k(S^n)$ is the standard $(n+1)$-dimensional space form of curvature k, i.e., an n-dimensional sphere of radius $1/\sqrt{k}$, a rescaled hyperbolic space, or $\mathbb{R}^n$, depending on the sign of k.

The primary use of cones in this chapter is to consider cones over spaces of directions. Note that a space X can be recovered as the space of directions of the cone $\mathrm{Con}_k(X)$ at the origin.

The results of Section 4.7 about curvature bounds of cones can be simplified for Alexandrov spaces in view of Toponogov's Theorem 10.3.1 (saying that the triangle comparison conditions hold in the large) and Corollary 10.4.2 (saying that the perimeter of any triangle in a space of curvature ≥ 1 is no greater than 2π).

Theorem 10.2.3. *Let X be a complete metric space and $k \in \mathbb{R}$.*

1. If X is an Alexandrov space of curvature ≥ 1, then $\mathrm{Con}_k(X)$ is an Alexandrov space of curvature $\geq k$.

2. If $\mathrm{Con}_k(X)$ is an Alexandrov space of curvature $\geq k$ and, in addition, $|xy| + |yz| + |xz| \leq 2\pi$ for all $x, y, z \in X$, then X is an Alexandrov space of curvature ≥ 1 or a space consisting of two points at distance π from each other.

The proof of this theorem is in no way different from that of Theorem 4.7.1. Note that in the second part of the theorem we do not assume in advance that X is a length space. Due to the other assumption, this property follows from the fact that $\mathrm{Con}_k(X)$ is a length space. Nor do we assume that X is connected—unlike the case of curvature bounded above, a cone over X cannot have a lower curvature bound if X is not connected (except the trivial case of a two-point space).

10.2.2. Quotients. Recall (see Section 3.3) that if a group Γ acts on a metric space (X, d) by isometries and the orbits $o(p) = \{\gamma p : \gamma \in \Gamma\}$ are closed, then we equip the quotient $Q = X/\Gamma$ with the strongest topology for

which the canonical projection $\pi : X \to Q$, $\pi(p) = o(p)$ is continuous. And we equip Q with the metric ρ such that

$$\rho(o(p), o(q)) = \inf\{d(p,r) | r \in o(q)\}. \tag{10.2}$$

Proposition 10.2.4. *If (X, d) is a length space of curvature $\geq k$ and a group Γ acts on X by isometries with closed orbits, then $Q = X/\Gamma$ also is a space of curvature $\geq k$.*

Proof. To simplify the considerations, we assume that X is locally compact. Let $p_0 \in Q$ and let $p \in \pi^{-1}(p_0)$ where $\pi : X \to Q$ is the projection map. Let $r > 0$ be such that the ball $U = B_r(p)$ is a normal region in the sense that the quadruple condition (cf. Proposition 10.1.1) is satisfied for any for quadruple contained in U. We will show that the same is true in the ball $U_0 = B_{r/2}(p_0)$ in the quotient space. Let $(a_0; b_0, c_0, d_0)$ be a quadruple in U_0. Since X is locally compact and the orbit $\pi^{-1}(a_0)$ is a closed set, there is a point $a \in \pi^{-1}(a_0)$ which is nearest to p, i.e., $|pa| = \operatorname{dist}(p, \pi^{-1}(a_0))$. The definition (10.2) of the quotient metric implies that $|pa| = |p_0a_0|$. Then one can similarly find points $b, c, d \in U$ such that $|ab| = |a_0b_0|$, $|ac| = |a_0c_0|$ and $|ad| = |a_0d_0|$. Note that $a, b, c, d \in U$. By (10.2), we have $|bc| \geq |b_0c_0|$, $|cd| \geq |c_0d_0|$ and $|db| \geq |d_0b_0|$; hence $\tilde{\angle} bac \geq \tilde{\angle} b_0a_0c_0$, $\tilde{\angle} cad \geq \tilde{\angle} c_0a_0d_0$ and $\tilde{\angle} bad \geq \tilde{\angle} b_0a_0d_0$. Therefore

$$\tilde{\angle} b_0a_0c_0 + \tilde{\angle} c_0a_0d_0 + \tilde{\angle} b_0a_0d_0 \leq \tilde{\angle} bac + \tilde{\angle} cad_0 + \tilde{\angle} bad \leq 2\pi$$

and the proposition follows. □

This proposition is commonly used to construct examples of spaces with curvature bounded below. One of the simplest examples of this kind is the following. Let $\mathbb{Z}_2 = \{e, \gamma\}$ act on $\mathbb{R}^3$ by symmetries: $\gamma(x) = -x$ for all $x \in \mathbb{R}^3$. Then $Q_2 = \mathbb{R}^3/\mathbb{Z}_2$ is a space of nonnegative curvature.

Exercise 10.2.5. Check that Q_2 is not a manifold. Prove that Q_2 is isometric to the cone $K(\mathbf{P}^2)$ over the projective space $\mathbf{P}^2$ equipped with the canonical metric of constant curvature 1.

Similarly, if a group G acts on the standard sphere $S^n \subset \mathbb{R}^{n+1}$ by isometries, then $Q = S^n/G$ is a space of curvature ≥ 1. Note that Q is a Riemannian manifold if G acts freely (i.e., without fixed points); otherwise Q may have metric or even topological singularities. As an example, consider the action of $G = Z_2 = \{e, \gamma\}$ on S^n defined by

$$\gamma(x_1, \dots, x_n, x_{n+1}) = (-x_1, \dots, -x_n, x_{n+1}).$$

Here the quotient space is the spherical cone over $\mathbf{P}^{n-1}$.

10.2.3. Convex surfaces in $\mathbb{R}^3$. Metrics of curvature bounded below arose initially in papers by A. D. Alexandrov as nonnegatively curved metrics on two-dimensional surfaces. Alexandrov proved that the class of two-dimensional nonnegatively curved length spaces essentially coincide with the class of convex surfaces in $\mathbb{R}^3$. In this section we prove one part of this coincidence, namely, that every convex surface in $\mathbb{R}^3$ (with its intrinsic metric) is a space of nonnegative curvature.

The other part is a remarkable theorem stating that every length space of nonnegative curvature homeomorphic to the 2-sphere is isometric to the boundary of a convex body in $\mathbb{R}^3$ or, as a degenerate case, to a twice covered planar convex region, i.e., two copies of a region glued together along the boundaries. From this theorem it easily follows that every point of a two-dimensional manifold with a metric of nonnegative curvature has a neighborhood isometric to a region on the boundary of a convex body in $\mathbb{R}^3$.

Similar results can be proved for metrics of curvature $\geq k$ and surfaces in spaces of constant curvature k. The restriction that the space is a manifold is not essential because one can prove that every two-dimensional space of curvature bounded below is a manifold (possibly with a boundary).

Theorem 10.2.6. *The intrinsic metric of a convex surface in $\mathbb{R}^3$ is a metric of nonnegative curvature.*

By a convex surface C we mean the boundary of a convex body X. It makes sense to add "degenerate surfaces" to make the class closed. The theorem is trivial for convex sets of dimensions 0 and 1. If $\dim X = 2$, we consider not the boundary of X but the twice covered region X, i.e. two copies of X glued together along their boundaries (convex curves). It is not difficult to prove the theorem in this case. The case of unbounded X is reduced to the compact case due to the local nature of the theorem.

Proof of Theorem 10.2.6. Now we assume that X is a convex body, that is, a compact convex set with nonempty interior. Let $C = \partial X$. The idea of the proof is very simple: we approximate X by convex polyhedra and prove that the intrinsic metrics of their boundaries converge to the intrinsic metric of C. It is easy to verify that the intrinsic metric of a convex polyhedral surface is nonnegatively curved. Later in Section 10.3 we will show that a Gromov–Hausdorff limit of nonnegatively curved spaces is itself nonnegatively curved (Proposition 10.7.1). Applying this fact finishes the proof. Now we pass to formal details.

Lemma 10.2.7. *If convex bodies $X_i \subset \mathbb{R}^3$ $(i \to \infty)$ converge in the Hausdorff metric to a convex body $X \subset \mathbb{R}^3$, then the intrinsic metrics of*

$C_i = \partial X_i$ converge uniformly (cf. Definition 7.1.5) to the intrinsic metric of $C = \partial X$.

Proof. We denote the intrinsic metrics of C and C_i by d and d_i respectively. We may assume that the origin 0 belongs to the interior of X and hence X contains a ball $B_r(0)$ for some $r > 0$. Since X is convex, every ray emanating from the origin intersects $C = \partial X$ exactly once. Hence the central projection from $\mathbb{R}^3$ to C (which maps every ray to its point of intersection with C) is well-defined. We will show that this central projection restricted to C_i is a homeomorphism from C_i to C (for all sufficiently large i) and its dilatation goes to zero as $i \to \infty$.

First consider arbitrary convex bodies X and X' in $\mathbb{R}^3$ such that $B_r(0) \subset X \subset X'$. Let $C = \partial X$, $C' = \partial X'$, d and d' denote the intrinsic metrics of C and C', and let δ be the maximum distance from a point in C' to its central projection in C. We are going to estimate how much the central projection from C' to C may enlarge the intrinsic distances. The key observation is the following "Busemann–Feller Lemma":

(i) For every point $a' \in \mathbb{R}^3$ there is a unique nearest point in X, i.e., a point $a \in X$ such that $|a - a'| = \operatorname{dist}(a', X)$.

(ii) If $a, b \in X$ are nearest points in X for $a', b' \in \mathbb{R}^3$, then $|a - b| \le |a' - b'|$.

To prove (i), suppose that a_1 and a_2 are two nearest points for a' in X. Then the midpoint $a = \frac{1}{2}(a_1 + a_2)$ belongs to X because X is convex. But $|a' - a| < \frac{1}{2}(|a' - a_1| + |a' - a_2|)$; hence a_1 and a_2 are not nearest points. To prove (ii), observe that the angles $\measuredangle abb'$ and $\measuredangle baa'$ are not less than $\pi/2$. Indeed, if $\measuredangle abb' < \pi/2$, then the segment $[ab] \subset X$ contains a point b_1 with $|b' - b_1| < |b' - b|$ (for example, pick b_1 close to b). Since the angles $\measuredangle abb' \ge \pi/2$ and $\measuredangle baa' \ge \pi/2$, the projection of the segment $[ab]$ to the line $a'b'$ contains the segment $[a'b']$. Hence $[ab]$ is not longer than $[a'b']$.

The Busemann–Feller Lemma can be reformulated as follows: the nearest-point map from $\mathbb{R}^3$ to X is well-defined and nonexpanding. We call this map the *orthogonal projection* to X. Since it is nonexpanding, it does not increase lengths of curves. Therefore if a', b' are two points in C' and a_0, b_0 are their orthogonal projections to X, then $d(a_0, b_0) \le d'(a', b')$. (Note that a_0 and b_0 belong to C because $X \subset X'$.)

Now let a and b be the central projections of a' and b' in C. Consider the straight segment $[aa']$ and project it onto C by means of orthogonal projection. The resulting curve γ connects a to a_0 in C, and $L(\gamma) \le L([aa']) \le \delta$. Hence $d(a, a_0) \le \delta$, and similarly $d(b, b_0) \le \delta$. It follows that $d(a, b) \le d'(a', b') + 2\delta$.

Now return to the original problem. We leave as an exercise to the reader the following fact about convex bodies: the Hausdorff convergence of $\{X_i\}$ to X implies that there is a sequence $\{\varepsilon_i\}$ of positive numbers such that $\varepsilon_i \to 0$ and $(1-\varepsilon_i)X \subset X_i \subset (1+\varepsilon_i)X$ for all large enough i. It follows that the central projection to X_i is well-defined. Applying the above estimate to the central projection from X_i to $(1-\varepsilon_i)X$, we obtain that

$$(1-\varepsilon_i)d(a,b) \le d_i(a_i,b_i) + 2\varepsilon_i D$$

where $D = \operatorname{diam}(X)$, a_i and b_i are central projections of a and b in X_i. Applying the same estimate to the central projection from $(1+\varepsilon_i)X$ to X_i, we obtain that

$$d(a_i,b_i) \le (1+\varepsilon_i)d(a,b) + 2\varepsilon_i D.$$

These two inequalities imply that the distortion of the central projection is not greater than $\varepsilon_i(\operatorname{diam}(C,d)+2D)$.

Since $\operatorname{diam}(C,d) < \infty$ (for example, $\operatorname{diam}(C,d) \le \pi D$ because the orthogonal projection from the sphere of radius D to C is nonexpanding and surjective), it follows that the distortions of central projections converge to zero. Hence (C_i,d_i) converge uniformly to (C,d). □

It is not difficult to approximate X by convex polyhedra P_i. For example, let P_i be a convex hull of a finite ε_i-net in X where $\varepsilon_i \to 0$. As we have seen in Chapter 4 (Theorem 4.2.14), the boundary of a convex polyhedron is a nonnegatively curved space. By the above lemma, C with its intrinsic metric is a uniform limit of these polyhedral spaces. Now the theorem follows from Proposition 10.7.1 that a lower curvature bound is preserved under Gromov–Hausdorff convergence. □

Generalizations. Theorem 10.2.6 holds for convex hypersurfaces in $\mathbb{R}^n$ for all $n \ge 3$. The proof is almost the same. (However, the converse theorem that every nonnegatively curved space is locally isometric to a convex surface is no longer true in higher dimensions.)

Moreover, one can replace $\mathbb{R}^n$ by a spherical or hyperbolic space of curvature k; in this case convex hypersurfaces have curvature $\ge k$. A more general (and not that easy) result that a convex hyper-surface in any Riemannian manifold of curvature $\ge k$ is itself a space of curvature $\ge k$, is proved by S. Buyalo.[1] It is still an open problem whether the boundary of a convex set in an Alexandrov space of curvature $\ge k$ is itself a space of curvature $\ge k$.

[1] S. Buyalo, *Shortest paths on convex hypersurfaces of a Riemannian manifold*, in Studies in Topology, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **66** (1976) 114-132 (Russian)

10.3. Toponogov's Theorem

Toponogov's Theorem is the "globalization theorem" for spaces of curvature bounded below. A space of curvature bounded below is defined in local terms; namely, certain triangle comparison inequalities are required to hold in a neighborhood of every point. Toponogov's Theorem says that this local definition implies that the same inequalities hold "in the large", i.e., for all triangles (cf. Definition 4.6.6 in Subsection 4.6.2). Unlike the case of curvature bounded above (Theorem 9.2.9), Toponogov's Theorem works for curvature bounds of any sign and, more importantly, does not require topological conditions like simple connectedness.

In the two-dimensional case this theorem was proved by A. Alexandrov. V. Toponogov proved it for Riemannian manifolds of any dimension. The general case is due to G. Perelman.

Theorem 10.3.1 (Toponogov's Theorem). *Let X be a complete length space of curvature $\geq k$. Then X has curvature $\geq k$ in the large.*

Remark 10.3.2. Recall that "in the large" means that the curvature comparison conditions are satisfied for all triangles for which comparison triangles exist and are unique.The latter requirement on triangles is essential only if $k > 0$ and is approximately equivalent to the statement that the perimeter is not greater than $2\pi/\sqrt{k}$. In Subsection 4.1.15 we will show that, in fact, an Alexandrov space of curvature $\geq k$ cannot contain a triangle of perimeter greater than $2\pi/\sqrt{k}$.

The proof of Theorem 10.3.1 is quite complicated and may be considered optional. To simplify the exposition, we prove the theorem only for locally compact spaces and leave to the reader most of the technical details arising in the case $k > 0$. Another proof of the theorem can be found in [**BGP**].

Proof of Theorem 10.3.1. The global versions of the definitions are equivalent just like the local ones. We will show that the "angle condition" 4.1.15 is satisfied for any triangle. Note that angles between shortest paths do exist and the sum of adjacent angles is equal to π because these properties are local.

We present a proof for the case $k \leq 0$ and then explain how to modify the argument for the case $k > 0$.

Step 1. Suppose that the theorem is false. Then there exists a triangle $\triangle pqr$ satisfying the following:

(i) The angle condition fails for its angle at q, i.e., $\angle pqr < \tilde{\angle} pqr$.

(ii) The angle condition is satisfied for any triangle of perimeter no greater than $0.99\,R$ whose vertices are contained in the ball $B_{100R}(q)$, where $R = \mathrm{per}(\triangle pqr) = |pq| + |pr| + |qr|$.

To prove the existence of such a triangle, suppose the contrary and construct a sequence of triangles $\{\triangle p_i q_i r_i\}_{i=1}^{\infty}$ as follows. Let $p_1q_1r_1$ be an arbitrary triangle such that $\measuredangle p_1q_1r_1 < \widetilde{\measuredangle} p_1q_1r_1$. Then for $i = 1, 2, \dots$ define $R_i = \mathrm{per}(\triangle p_iq_ir_i)$ and let $\triangle p_{i+1}q_{i+1}r_{i+1}$ be a triangle of perimeter no greater than $0.99\,R_i$, with vertices in $B_{100R_i}(q_i)$, and such that

$$\measuredangle p_{i+1}q_{i+1}r_{i+1} < \widetilde{\measuredangle} p_{i+1}q_{i+1}r_{i+1}.$$

Then $R_{i+1} \le 0.99\,R_i \le \big(0.99\big)^i R_1$. Since $|q_iq_{i+1}| \le 100R_i$, it follows that the series $\sum_{i=1}^{\infty} |q_iq_{i+1}|$ converges; hence $\{q_i\}$ is a Cauchy sequence. Let $q = \lim_{i\to\infty} q_i$ and U be a normal neighborhood of p. Then $\triangle p_iq_ir_i \subset U$ for a large enough i; hence $\triangle p_iq_ir_i$ must satisfy the angle condition. This is a contradiction.

Step 2. Let $\triangle pqr$ satisfy the conditions (i) and (ii) from Step 1. Then there exists a triangle $\triangle abc$ with vertices in $B_R(q)$, of perimeter not greater than R, which is "thin" in the sense that $|ac| < 0.01\,R$, and such that the angle condition fails for $\measuredangle bac$.

To prove this, divide the side $[qr]$ of $\triangle pqr$ into small intervals by points $c_0 = q$, c_1, $\dots$, $c_n = r$ so that each triangle $\triangle c_ipc_{i+1}$ is "thin" ($|c_ic_{i+1}| < 0.01\,R$). Note that the perimeters of the triangles $\triangle c_ipc_{i+1}$ are not greater than $\mathrm{per}(\triangle pqr) \le R$. We will show that the angle condition fails for $\measuredangle pc_ic_{i+1}$ or $\measuredangle pc_{i+1}c_i$ in one of these triangles. Suppose the contrary. Then it follows by induction that the angle condition is satisfied for the angles at q and c_i in $\triangle qpc_i$ for $i = 1, \dots, n$. Indeed, assuming this for $\triangle qpc_i$ ($i < n$), place comparison triangles $\triangle \bar{q}\bar{p}\bar{c}_i$ and $\triangle \bar{c}_i\bar{p}\bar{c}_{i+1}$ in different half-planes with respect to their common side $\bar{p}\bar{c}_i$ in the k-plane. Then Lemma 4.3.3 implies that

$$\widetilde{\measuredangle} pqc_{i+1} \le \measuredangle \bar{p}\bar{q}\bar{c}_i = \widetilde{\measuredangle} pqc_i \le \measuredangle pqc_i = \measuredangle pqc_{i+1}$$

and similarly $\widetilde{\measuredangle} pc_{i+1}q \le \measuredangle pc_{i+1}q$. This completes the induction step. Now let $i = n$; then the statement we just proved implies that the angle condition is satisfied for $\measuredangle pqr$, contrary to our assumptions.

Step 3. Let $\triangle abc$ be a "thin" triangle constructed in Step 2, and let d be the midpoint of $[bc]$. Then $|bd| = |cd| \le \frac{1}{4}R$, $|ad| \le 0.26\,R$ and $\mathrm{per}(\triangle adc) \le 0.52\,R$; hence the angle condition holds for the angles of $\triangle adc$. Therefore by Lemma 4.3.3 (compare with Step 2), it follows that $\measuredangle adb < \widetilde{\measuredangle} adb$. Choose an $\varepsilon > 0$ such that

$$\measuredangle adb \le \widetilde{\measuredangle} adb - \varepsilon = \measuredangle \bar{a}\bar{d}\bar{b} - \varepsilon \qquad (10.3)$$

where $\triangle \bar{a}\bar{d}\bar{b}$ is a comparison triangle.

Let us take a point x close to d and place comparison triangles for $\triangle adb$, $\triangle axb$, $\triangle axd$, and $\triangle dxb$ as shown in Figure 10.3. We want a point x to be such that the comparison triangles $\triangle \bar{a}\bar{b}\bar{x}_3$, $\triangle \bar{d}\bar{b}\bar{x}_2$ and $\triangle \bar{a}\bar{d}\bar{x}_1$ have disjoint interiors and $\measuredangle \bar{x}_1\bar{d}\bar{x}_2 \geq \varepsilon$. To obtain such an x we do the following: take a_1 and b_1 close to d in the shortest paths $[ad]$ and $[bd]$, and let x be the midpoint of a shortest path $[a_1b_1]$.

Observe that the angle condition is satisfied for triangles $\triangle aa_1x$, $\triangle a_1dx$, $\triangle bb_1x$, $\triangle b_1dx$ and $\triangle a_1db_1$ because their perimeters are smaller than $0.99\,R$. Using this and the implied monotonicity of angles, we obtain

$$\measuredangle \bar{a}\bar{d}\bar{x}_1 + \measuredangle \bar{b}\bar{d}\bar{x}_2 \leq \widetilde{\measuredangle} a_1dx_1 + \widetilde{\measuredangle} b_1dx_2 \leq \widetilde{\measuredangle} a_1db_1 \leq \measuredangle adb \leq \measuredangle \bar{a}\bar{d}\bar{b} - \varepsilon.$$

Hence the triangles $\triangle \bar{a}\bar{d}\bar{x}_1$, and $\triangle \bar{b}\bar{d}\bar{x}_2$ in Figure 10.3 have no common interior points and

$$\measuredangle \bar{x}_1\bar{d}\bar{x}_2 = \measuredangle \bar{a}\bar{d}\bar{b} - \measuredangle \bar{a}\bar{d}\bar{x}_1 - \measuredangle \bar{b}\bar{d}\bar{x}_2 \geq \varepsilon.$$

Since $|\bar{a}\bar{x}_3| = |\bar{a}\bar{x}_1|$ and $|\bar{b}\bar{x}_3| = |\bar{b}\bar{x}_2|$, it then follows that $\triangle \bar{a}\bar{b}\bar{x}_3$ has no common interior points with $\triangle \bar{a}\bar{d}\bar{x}_1$, and $\triangle \bar{b}\bar{d}\bar{x}_2$.

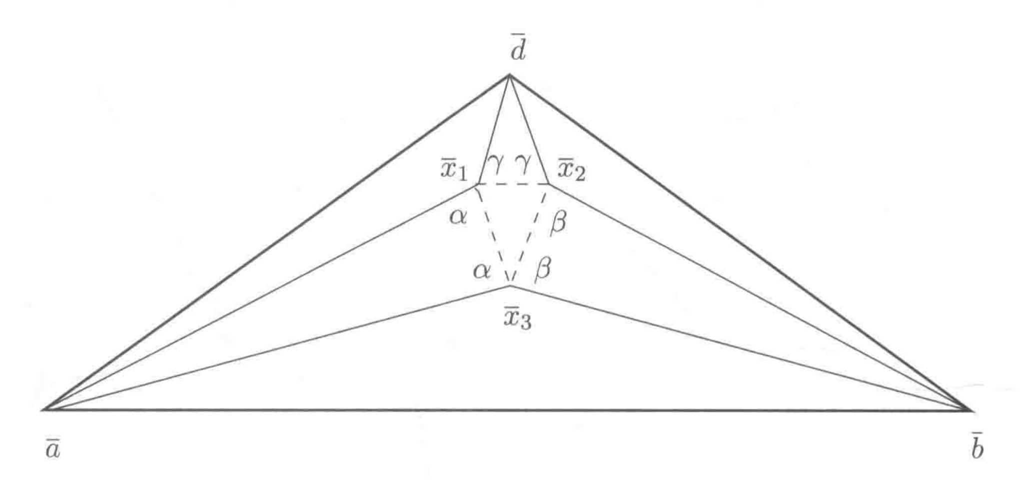


Figure 10.1: Toponogov's theorem, Step 3.

Step 4. Let $\alpha = \measuredangle \bar{a}\bar{x}_1\bar{x}_3 = \measuredangle \bar{a}\bar{x}_3\bar{x}_1$, $\beta = \measuredangle \bar{b}\bar{x}_2\bar{x}_3 = \measuredangle \bar{b}\bar{x}_3\bar{x}_2$ and $\gamma = \measuredangle \bar{d}\bar{x}_1\bar{x}_2 = \measuredangle \bar{d}\bar{x}_2\bar{x}_1$ (see Figure 10.3). Now, applying the Gauss–Bonnet formula to the triangle $\triangle \bar{x}_1\bar{x}_2\bar{x}_3$ and taking into account that $2\alpha \leq \pi$, $2\beta \leq \pi$, $2\gamma + \varepsilon \leq 2\gamma + \measuredangle \bar{x}_1\bar{d}\bar{x}_2 \leq \pi$ (here we use our assumption that $k \leq 0$), we obtain that

$$\begin{aligned} 2\pi &= \omega(\triangle \bar{x}_1\bar{x}_2\bar{x}_3) + 2(\alpha+\beta+\gamma) + (\measuredangle \bar{a}\bar{x}_1\bar{d} + \measuredangle \bar{d}\bar{x}_2\bar{b} + \measuredangle \bar{a}\bar{x}_3\bar{b}) - 3\pi \\ &\leq (3\pi - \varepsilon) + (\measuredangle axd + \measuredangle dxb + \measuredangle axb) + (-\measuredangle axb + \measuredangle \bar{a}\bar{x}_3\bar{b}) - 3\pi \end{aligned}$$

where ω is the integral curvature, $\omega(E) = k \cdot \text{Area}(E) \le 0$. Since $\measuredangle axd + \measuredangle dxb + \measuredangle axb \le 2\pi$, it follows that

$$\widetilde{\measuredangle} axb - \measuredangle axb = \measuredangle \bar{a}\bar{x}_3\bar{b} - \measuredangle axb > \varepsilon.$$

On the other hand, $|ax| + |bx| \le |ad| + |db| - \frac{\varepsilon}{10}|dx|$ if $|dx|$ is small enough.

Step 5. Consider the set S of all triangles $\triangle ayb$ (with a and b fixed) such that

(i) $\widetilde{\measuredangle} ayb - \measuredangle ayb \ge \varepsilon$, and

(ii) $\max\{|ay|, |by|\} \le 0.26\, R$.

We supposed that this set was nonempty. Since our space is locally compact, there is a triangle $\triangle abd \in S$ with

$$|ad| + |bd| = \min\{|ay| + |yb| : \triangle ayb \in S\}.$$

However Step 4 shows that one can find a point x (close to d) such that $\triangle axb \in S$ and $|ax| + |xb| < |ad| + |db|$. This contradiction proves the theorem for the case $k \le 0$.

Step 6. If $k > 0$, the same scheme works with certain modifications. First, one has to check carefully that the involved comparison triangles are well-defined. Second, in the inequalities that follow from the Gauss–Bonnet formula in Step 4, one cannot drop the integral curvature of the triangles $\triangle \bar{x}_1\bar{x}_2\bar{x}_3$ and $\bar{d}\bar{x}_1\bar{x}_2$. As a result, one loses the condition (i) in Step 5 when $\triangle adb$ is replaced by $\triangle axb$. To work this around, replace the condition (i) in Step 5 by a more complicated one, namely,

$$(\widetilde{\measuredangle} ayb - \measuredangle ayb) - c(|ay| + |yb|) \ge \frac{1}{2}\varepsilon, \tag{10.4}$$

where c is a sufficiently small positive constant. Then the loss in the first term of (10.4) is compensated by the gain in the second one. The details are left to the reader as an exercise .

The reader who has checked all these details might have noticed that the argument works only for triangles of perimeter strictly less than $2\pi/\sqrt{k}$. So we have to work out the case of a triangle $\triangle abc$ with perimeter $|ab| + |bc| + |ac| = 2\pi/\sqrt{k}$, and prove that $\measuredangle abc = \pi$ if $\max\{|ab|, |bc|\} < \pi/\sqrt{k}$. Fix a positive $\varepsilon < k$. Observe that a space of curvature $\ge k$ is also a space of curvature $k - \varepsilon$. Applying the already proven part of the theorem to X as a space of curvature $\ge k - \varepsilon$, we obtain that $\measuredangle abc \ge \widetilde{\measuredangle}_{k-\varepsilon} abc$. As $\varepsilon \to 0$, a comparison triangle for $\triangle abc$ in the $(k-\varepsilon)$-plane converges to a comparison triangle in the k-plane. Therefore $\widetilde{\measuredangle}_{k-\varepsilon} abc \to \widetilde{\measuredangle}_k abc = \pi$ as $\varepsilon \to 0$, implying that $\measuredangle abc \ge \pi$. $\square$

10.4. Curvature and Diameter

This section can be viewed as an addendum to Toponogov's Theorem in the case of positive curvature bound k. The conclusion of the theorem reads: if a triangle is "not too large" so that it has a well-defined (i.e., unique up to an isometry) comparison triangle in the k-plane, then it possesses the comparison properties. "Not too large" here means that the perimeter is not greater than $2\pi/\sqrt{k}$. (One should also exclude triangles $\triangle abc$ with $|ac| = \pi/\sqrt{k}$ and $|ab| + |bc| = \pi/\sqrt{k}$ for which a comparison triangle is not unique. Or, alternatively, choose comparison triangles such that the side corresponding to $[ac]$ passes through the point corresponding to b.)

A question arises: what about triangles whose perimeters are greater than $2\pi/\sqrt{k}$? In this section we give an easy answer: such triangles simply do not exist. Recall that we have excluded some spaces from the class of Alexandrov spaces of curvature $\geq k$. These exceptions are: circles of length greater than $2\pi/\sqrt{k}$, line segments of length greater than $\pi/\sqrt{k}$, the half-line $\mathbb{R}_+$, and the line $\mathbb{R}$.

Theorem 10.4.1. *Let X be an Alexandrov space of curvature $\geq k$ where $k > 0$. Then* $\operatorname{diam}(X) \leq \pi/\sqrt{k}$.

Proof. Suppose the contrary, and let $a, b \in X$ be two points such that $|ab| > \pi/\sqrt{k}$. We may assume that $|ab| = (\pi+\varepsilon)/\sqrt{k}$ where $0 < \varepsilon < \pi/4$. Let c be the midpoint of a shortest path $[ab]$ and U be an $(\varepsilon/3\sqrt{k})$-neighborhood of c.

First we show that U contains a point which does not belong to $[ab]$. Indeed, suppose the contrary. For every point $x \in X$ consider a shortest path γ connecting c to x. Our assumption implies that this shortest path coincides with a subinterval of $[ab]$ in a neighborhood of c. Since geodesics do not branch (Exercise 10.1.2), it follows that x belongs to the unique geodesic containing $[ac]$. Therefore X is covered by two minimal geodesics (that is, geodesics whose intervals are shortest paths) starting at c and passing through a and b respectively. Depending on whether these geodesics are finite or infinite and whether their endpoints coincide, we conclude that X is one of the one-dimensional exceptional spaces.

Choose an $x \in U \setminus [ab]$, and let y be a nearest point to x in $[ab]$. Then $|ay| > \pi/2\sqrt{k}$ and $|by| > \pi/2\sqrt{k}$. The first variation formula implies (see Corollary 4.5.7) that $\angle xya = \angle xyb = \pi/2$. Consider a comparison triangle $\triangle\bar{x}\bar{y}\bar{a}$ for $\triangle xya$ in the k-plane which is the sphere of radius $1/\sqrt{k}$. By Toponogov's Theorem we have $\angle\bar{x}\bar{y}\bar{a} \leq \angle xya = \pi/2$. Since $|\bar{y}\bar{a}| > \pi/2\sqrt{k}$, it follows that $|\bar{x}\bar{a}| < |\bar{y}\bar{a}|$. (To see this, place $\bar{a}$ at the north pole of the sphere; then the inequality $\angle\bar{x}\bar{y}\bar{a} \leq \pi/2$ implies that $\bar{x}$ lies above the great circle passing through $\bar{y}$ orthogonally to the meridian $[\bar{a}\bar{x}]$. Since $\bar{y}$ is strictly

below the equator, $\bar{y}$ is the lowest point of this great circle; hence $\bar{y}$ is strictly lower than $\bar{x}$, i.e., $|\bar{x}\bar{a}| < |\bar{y}\bar{a}|$.) Thus $|xa| < |ya|$ and similarly $|xb| < |yb|$. Hence $|ya| + |yb| > |xa| + |xb| \geq |ab|$, contrary to the fact y belongs to a shortest path $[ab]$. □

Corollary 10.4.2. *Let X be an Alexandrov space of curvature $\geq k$ where $k > 0$. Then any triangle in X has perimeter no greater than $2\pi/\sqrt{k}$.*

Proof. We first prove the theorem under an additional assumption that $\operatorname{diam}(X)$ is strictly less than $\pi/\sqrt{k}$. Suppose the contrary and let $x, y, z \in X$ be such that $|xy| + |yz| + |xz| > 2\pi/\sqrt{k}$. Fix shortest paths $[xy]$ and $[xz]$. By continuity, there exist points $y' \in [xy]$ and $z' \in [xz]$ such that $|xy'| + |xz'| + |y'z'| = 2\pi/\sqrt{k}$. Consider a triangle $\triangle xy'z'$. Since $\operatorname{diam}(X) < \pi/\sqrt{k}$, its sides are shorter than $\pi/\sqrt{k}$; hence a comparison triangle $\triangle\bar{x}\bar{y}'\bar{z}'$ for $\triangle xy'z'$ in the k-plane is well-defined. Since the perimeter of $\triangle\bar{x}\bar{y}'\bar{z}'$ equals $2\pi/\sqrt{k}$, its angles equal π; therefore $\measuredangle xy'z' = \measuredangle xz'y' = \pi$ by Toponogov's Theorem. Hence $\measuredangle zy'z' = \measuredangle yz'y' = 0$, and this implies that the shortest path $[y'z']$ passes through y and z. Then the perimeter of $\triangle xyz$ is less or equal to that of $\triangle xy'z'$. This contradiction proves the statement in the case $\operatorname{diam}(X) < \pi/\sqrt{k}$.

Now we pass to the general case. Fix a positive $\varepsilon < k$. By Theorem 10.4.1 we have $\operatorname{diam}(X) \leq \pi/\sqrt{k} < \pi/\sqrt{k-\varepsilon}$. Furthermore, since X is a space of curvature $\geq k$, it is also a space of curvature $\geq k - \varepsilon$. Therefore we can apply the above arguments with k replaced by $k - \varepsilon$ and conclude that the perimeter of any triangle is no greater than $2\pi/\sqrt{k-\varepsilon}$. Since ε is arbitrary, the desired statement follows. □

Exercise 10.4.3. Let X be a compact space of curvature ≥ 1 and $\operatorname{diam}(X) = \pi$. Prove that X is isometric to a suspension (i.e., spherical cone; cf. Subsection 10.2.1) over a compact space of curvature ≥ 1.

Hint: Let $p, q \in X$ be such that $|pq| = \pi$. Define

$$Y = \{x \in X : |px| = |qx| = \pi/2\}.$$

Then prove the following facts. First, for any $x, y \in Y$, shortest paths $[px]$ and $[qx]$ are unique and $\measuredangle pxy = \measuredangle qxy = \measuredangle pyx = \measuredangle qyx = \pi/2$. Then the triangles $\triangle pxy$ and $\triangle qxy$ (as unions of shortest paths) are isometric to their comparison triangles in the sphere. Then Y is convex in X and therefore Y is a length space of curvature ≥ 1. Finally, every point $x \in X$ belongs to some shortest path connecting p and q, and therefore to the union $[py] \cup [qy]$ for some $y \in Y$. These facts imply that X is the spherical cone over Y.

Another proof is based on the splitting theorem from the next section; see Remark 10.5.8.

Exercise 10.4.4 (Berger's theorem). Let X be an n-dimensional Riemannian manifold of curvature $\geq k$ and $\operatorname{diam}(X) = \pi$. Prove that X is isometric to the standard sphere S^n.

Exercise 10.4.5. The *radius* $\operatorname{rad}(X)$ of a compact metric space X is the maximal number $r > 0$ such that $X = \overline{B}_r(p)$ for some $p \in X$. Prove that

1. $\frac{1}{2}\operatorname{diam}(X) \leq \operatorname{rad}(X) \leq \operatorname{diam}(X)$ for any metric space X.

2. If X is an n-dimensional Alexandrov space of curvature ≥ 1 and $\operatorname{rad}(X) = \pi$, then X is isometric to S^n.

Hint: Using the statement of Exercise 10.4.3, prove that X contains subsets isometric to S^1, S^2, ...

10.5. Splitting Theorem

Recall that a geodesic $\gamma : (-\infty, \infty) \to X$ is said to be a straight line (or, briefly, a line) if every one of its segments is a shortest path between its endpoints. Here is a remarkable splitting theorem.

Theorem 10.5.1 (Splitting theorem). *If a locally compact Alexandrov space X of nonnegative curvature contains a line, then X is isometric to a direct metric product $\mathbb{R} \times Y$ for some nonnegatively curved Alexandrov space Y.*

This theorem was proved by V. Toponogov for Riemannian manifolds and generalized by A. Milka for general spaces of nonnegative curvature.[2]

Remark 10.5.2. For Riemannian manifolds, a stronger theorem holds: one can assume nonnegativity of Ricci curvatures instead of nonnegativity of sectional curvatures (see [**CG**]).

In the course of the proof of the theorem, X always denotes an Alexandrov space of nonnegative curvature. We introduce the following definition of parallel lines (used in this section only and having no relation to parallel lines considered in Chapter 9).

Definition 10.5.3. Two straight lines γ, γ_1 in X are said to be *parallel* if there are two straight lines $\overline{\gamma}$, $\overline{\gamma}_1$ in $\mathbb{R}^2$ and an isometry of the metric space $\gamma \cup \gamma_1$ onto the metric space $\overline{\gamma} \cup \overline{\gamma}_1$.

Here we mean an isometry with respect to the restrictions of the metrics of X and $\mathbb{R}^2$ to these sets.

We say that the lines $\overline{\gamma}$ and $\overline{\gamma}_1$ in $\mathbb{R}^2$ *correspond* to the lines γ and γ_1 in X. We fix an isometry from $\gamma \cup \gamma_1$ to $\overline{\gamma} \cup \overline{\gamma}_1$ and refer to the image of a point $a \in \gamma \cup \gamma_1$ under this isometry as the point *corresponding to a*.

---

[2] A. D. Milka, *Metric structure of some class of spaces containing straight lines*, Ukrain. Geometrical. Sbornik, vyp. 4, 1967, Kharkov, pp. 43-48 (in Russian). In dim 2 this theorem was proved by Cohn-Vossen much earlier.

Lemma 10.5.4. *Let γ be a line in X and $p \in X$. Then for every two points $x, y \in \gamma$ the angles at x and y of a triangle $\triangle pxy$ are equal to the respective angles of its comparison triangle $\widetilde{\triangle} pxy$.*

Proof. Suppose that $\widetilde{\measuredangle} pxy \leq \measuredangle pxy - \varepsilon$, $\varepsilon > 0$. Take two points $z_i \in \gamma$ such that x is located between z_1 and z_2, and y is between z_1 and x. Place comparison triangles $\widetilde{\triangle} pxz_1$ and $\widetilde{\triangle} pxz_2$ in different half-planes with respect to their common side $\bar{p}\bar{x}$. And place a comparison triangle $\triangle \bar{p}\bar{x}\bar{y}$ in the same half-plane where the triangle $\widetilde{\triangle} pxz_1$ is placed. The rectangles $\bar{p}\bar{z}_1\bar{x}\bar{z}_2$ and $\bar{p}\bar{y}\bar{x}\bar{z}_2$ are convex. Due to the angle monotonicity definition we have $\widetilde{\measuredangle} pxz_1 \leq \widetilde{\measuredangle} pxy \leq \measuredangle pxy - \varepsilon$ and, therefore, the angle $\widetilde{\measuredangle} z_1xz_2 < \pi - \varepsilon$. This is impossible if z_1 and z_2 are sufficiently far from x because $|z_1p| + |pz_2| \geq |z_1x| + |xz_2|$. $\square$

By monotonicity of angles the lemma immediately implies the following

Corollary 10.5.5. *1. For points $p' \in [px]$ and $y' \in [yx]$ and points $\bar{p}' \in [\bar{p}\bar{x}]$ and $\bar{y}' \in [\bar{y}\bar{x}]$ such that $|p'x| = |\bar{p}'\bar{x}|$, $|y'x| = |\bar{y}'\bar{x}|$ the equality $|p'y'| = |\bar{p}'\bar{y}'|$ holds.*

2. For a line γ and a point p there is an unique nearest point to p in γ.

This nearest point is called the *projection* of p to γ. (Note however that a shortest path $[pp']$ may not be unique.)

Lemma 10.5.6. *1. For every line γ and every point p, there is no more than one line γ_1 parallel to γ and passing through p.*

2. The relation of being parallel is transitive.

Proof. Suppose that there are two lines, γ_1 and γ_2, passing through p and parallel to γ. Consider three corresponding parallel lines $\bar{\gamma}_1$, $\bar{\gamma}$, and $\bar{\gamma}_2$ in the plane $\mathbb{R}^2$. We assume that $\bar{\gamma}$ lies between the two other lines. Let points $\bar{a} \in \bar{\gamma}_1$, $\bar{b} \in \bar{\gamma}_2$ correspond to points $a \in \gamma_1$, $b \in \gamma_2$ such that $|pa| = |pb|$ and let the segment $[\bar{a}\bar{b}]$ intersect the line $\bar{\gamma}$ at point $\bar{c}$. Then $|ab| \leq |ac| + cb| = |\bar{a}\bar{c}| + |\bar{a}c\bar{b}| = |\bar{a}\bar{c}|$. Applying Lemma 10.5.4 to the point a and the line γ_2, one sees that $|ab| \to \infty$ as $|pa| = |pb| \to \infty$. The latter is impossible because $|\bar{a}\bar{c}|$ is a constant.

The second assertion of the lemma is proved similarly. The only difference is that one should consider four parallel lines in the plane. $\square$

Proof of the theorem. *Step 1.* Let γ be a straight line. First of all, for every point $p \in X \setminus \gamma$, we construct a straight line γ_p passing through p and parallel to γ. To do this, denote $c_i = \gamma(t_i)$ and consider shortest paths $[pc_i]$ where $t_i \to \infty$ as $i \to \infty$. There is a subsequence of these paths converging to a geodesic ray $\gamma_p^+ : [0, \infty) \to X$. Similarly, taking points $d_i = \gamma(-t_i)$,

$t_i \to \infty$, one can find a geodesic ray $\gamma_p^- : [0, -\infty)) \to X$. We are going to show that these two rays together form a straight line γ_p parallel to γ.

The comparison angles $\widetilde{\measuredangle} c_i p d_i$ obviously approach π as $i \to \infty$. Since $\measuredangle c_i p d_i \geq \widetilde{\measuredangle} c_i p d_i$, this implies that angles $\measuredangle c_i p d_i$ converge to π as well. Hence the angle between γ_p^+ and γ_p^- at p equals π.

Now let us prove that γ_p is a line. Fix a $t > 0$ and let b_i^+ and b_i^- be the points in the shortest paths $[pc_i]$ and $[pd_i]$ at the distance t from p. These points converge to the points $b^+ = \gamma_p(t)$, $b^- = \gamma_p(-t)$. Consider a comparison triangle $\triangle \bar{c}_i \bar{a} \bar{d}_i$ and points $\bar{b}_i^+$, $\bar{b}_i^-$ in its sides at distances t of $\bar{a}$. It is obvious that $|b^+ b^-| = \lim_{i\to\infty} |b_i^+ b_i^-| \geq \lim_{i\to\infty} |\bar{b}_i^+ \bar{b}_i^-| = 2t$. This means that each segment of γ_p is a shortest path; hence γ_p is a line.

Now we prove that γ and γ_p are parallel. Consider two lines $\bar{\gamma}$, $\overline{\gamma_p}$ at the distance $|pp'|$ in a plane $\mathbb{R}^2$ where p' is the projection of p to γ. Let $\bar{p} \in \bar{\gamma}_p$ and $\bar{p}'$ be the projection of $\bar{p}$ to $\bar{\gamma}$. Then the limit process used above and Corollary 10.5.5 imply that for every two points $a \in \gamma$, $b \in \gamma_a$ and points $\bar{a} \in \bar{\gamma}$, $\bar{b} \in \overline{\gamma_a}$, lying in the corresponding rays and such that $|pa| = |\bar{p}\bar{a}|$, $|p'b| = |\bar{p}'\bar{b}|$, the equality $|ab| = |\bar{a}\bar{b}|$ holds. This means that lines γ and γ_p are parallel.

Note that we have proved that for any sequence $\{t_i\}$, $t_i \to +\infty$, shortest paths $[p\gamma(t_i)]$ converge to a half-line of a geodesic parallel to γ if they converge at all. Since such a geodesic is unique (Lemma 10.5.6), it follows that these shortest paths converge to this half-line γ_p^+ for any sequence $\{t_i\}$, $t_i \to +\infty$. Moreover the angles between these paths and γ_p^+ go to zero.

Step 2. Applying Lemma 10.5.6, we see now that X is split into parallel lines, and we have the relation "one is the projection of the other" between points. We are going to show that this relation is symmetric and transitive.

To prove that this relation is symmetric, consider two parallel lines γ_1 and γ_2 and points $a_1 \in \gamma_1$, $a_2 \in \gamma_2$ such that a_2 is the projection of a_1 to γ_2. Consider triangles $a_1 a_2 \gamma(t)$ as $|t| \to \infty$. Since $\measuredangle a_1 a_2 \gamma(t) = \widetilde{\measuredangle} a_1 a_2 \gamma_2(t) = \pi/2$ and $|a_2 \gamma_2(t)| \to \infty$, we have $\measuredangle a_2 a_1 \gamma_2(t) = \widetilde{\measuredangle} a_2 a_1 \gamma_2(t) \to \pi/2$. Therefore $[a_1 a_2]$ is orthogonal to γ_1, or, equivalently, a_1 is the projection of a_2 to γ_1.

To prove that the projection relation is transitive, consider three parallel lines γ_1, γ_2, γ_3 and let a_2 and a_3 be the projections of a point $a_1 \in \gamma_1$ to γ_2 and γ_3, respectively. Then

$$|a_2 \gamma_1(t)| - |a_3 \gamma_1(t)| \to 0, \qquad t \to +\infty,$$

because $|a_2\gamma_1(t)| - |a_1\gamma_1(t)| \to 0$ and $|a_3\gamma_1(t)| - |a_1\gamma_1(t)| \to 0$. It follows that $\widetilde{\measuredangle} a_2 a_3 \gamma(t) \to \pi/2$, and therefore $[a_2 a_3]$ is orthogonal to $\gamma_3^+ = \lim[a_3, \gamma_1(t)]$. Hence a_3 is the projection of a_2 to γ_3.

Step 3. Now it is not difficult to finish the proof of the theorem. For a point $p \in \gamma$ denote by $N(p)$ the set of projections of p to all lines parallel to γ. This set is convex in X and hence is a length space of nonnegative curvature. Every two such sets $N(p)$ and $N(q)$ are canonically isometric via an isometry defined as follows: if $x \in N(p)$, then the image of x is $N(q) \cap \gamma_q$, where γ_q is the line parallel to γ and passing through q.

Fixing $p = \gamma(0)$, consider the set $N(p) \times \mathbb{R}^1$ and the map $f : N(p) \times \mathbb{R}^1 \to M$ such that $f(x, t)$ is the point of $N(\gamma(t))$ which corresponds to x under the canonical isometry between $N(p)$ and $N(\gamma(t))$. It is clear that f is an isometry. □

Remark 10.5.7. There is a different (and ideologically somewhat better) exposition of the proof of the theorem based on convexity properties of Busemann functions and horoballs (see 5.3.3). In Hadamard spaces (outward) equidistants (r-neighborhoods) of convex sets are convex, and hence so are all horoballs. Contrary to that, in spaces of nonnegative curvature inward equidistants preserve convexity; in particular, the reader can prove as an exercise that complements to horoballs are convex. This allows us to construct a family of convex sets exhausting X. In the case of manifolds, one can use smoothness properties of horoballs. Dividing the line (from the condition of the theorem) into two rays and studying their Busemann functions and corresponding families of horoballs easily gives a proof of a Riemannian case of the theorem. A similar argument allows us to prove the Cheeger–Gromoll Soul Theorem. For Alexandrov spaces, where even defining boundaries of convex sets requires a hard Stratification Theorem, this approach encounters technical difficulties. This is the reason why we use a modification of the proof (following an argument due to Milka) avoiding such difficulties.

Remark 10.5.8. The Splitting Theorem implies that an Alexandrov space X of curvature ≥ 1 and with diameter π is a spherical cone (see Exercise 10.4.3). Indeed, let Y be the cone over X; then Y contains a line (corresponding to a pair of points in X at the distance π from each other). Then Y is a direct product of $\mathbb{R}$ and some nonnegatively curved space Z. Since X is a space of directions of Y ($X = \Sigma_o(Y)$ where o is the origin), it follows that X is a spherical cone over the space of directions $\Sigma_o(Z)$.

10.6. Dimension and Volume

10.6.1. Dimensional homogeneity.

Theorem 10.6.1. *Let X be a complete locally compact space of curvature bounded below. Then all open subsets of X have the same Hausdorff dimension, namely $\dim_H(U) = \dim_H(X)$ for any open set $U \subset X$.*

The property expressed in the theorem is referred to as *dimensional homogeneity* of X. We do not exclude *a priori* noninteger and infinite values for dimension.

Proof. Every open set U contains a ball $B_r(p)$ for some $p \in X$ and $r > 0$, and $\dim_H(B_r(p)) \le \dim_H(U) \le \dim_H(X)$ because $B_r(p) \subset U \subset X$. It is sufficient to prove that $\dim_H(B_r(p)) = \dim_H(X)$. We will show that $\dim_H(B_r(p)) = \dim_H(B_R(p))$ for any $R > r$. The theorem follows from this, because X is can be represented as a union of countably many balls of the form $B_R(p)$; for example, take $R = r+1, r+2, \dots$. If the dimensions of these balls equal $\dim_H(B_r(p))$, it follows that $\dim_H(X) = \dim_H(B_r(p))$; cf. Proposition 1.7.19.

Denote $A = B_r(p)$ and $B = B_R(p)$. The inequality $\dim_H(A) \le \dim_H(B)$ is trivial because $A \subset B$. To prove that $\dim_H(B) \le \dim_H(A)$, we construct a map $f : B \to A$ such that for some $c > 0$ one has $|f(x)f(y)| \ge c|xy|$ for all $x, y \in A$. If such a map f exists, its inverse $f^{-1} : f(B) \to B$ is Lipschitz; therefore by Proposition 1.7.19 we have

$$\dim_H(B) \le \dim_H(f(B)) \le \dim_H(A).$$

This map f is used so many times in this chapter that it deserves a special description:

Lemma 10.6.2. *Let X be a complete locally compact space of curvature $\ge k$, $p \in X$, and $0 < \lambda < 1$. For every $x \in X$, let $f(x)$ be a point in some (arbitrarily chosen) shortest path $[px]$ such that $|pf(x)| = \lambda|px|$. Then*

(1) *If $k = 0$, then $|f(x)f(y)| \ge \lambda \cdot |xy|$ for all $x, y \in X$.*
(2) *If $k < 0$, then for any $R > 0$ there is a positive number $c(k, \lambda, R)$ such that $|f(x)f(y)| \ge c(k, \lambda, R) \cdot |xy|$ for all $x, y \in B_R(p)$.*

In fact, one can take $c(k, \lambda, R) = \frac{\sinh(-k\lambda R)}{\sinh(-kR)}$.

Proof. Follows immediately from monotonicity of angles. The number $c(k, \lambda, R)$ is a coefficient for which the inequality holds for the similar map in an R-ball in the k-plane. (The map in the k-plane is a diffeomorphism and hence is bi-Lipschitz in any ball, so a desired constant exists. A straightforward computation yields $c(k, \lambda, R) = \frac{\sinh(-k\lambda R)}{\sinh(-kR)}$.) □

We refer to the map f from the lemma as a *λ-homothety centered at p.*

To finish the proof of the theorem, set $\lambda = r/R$ and let f be a λ-homothety centered at p. Then $f(B) = f(B_R(p)) \subset B_r(p) = p$ and

$$|f(x)f(y)| \ge c(k, \lambda, R) \cdot |xy|$$

for all $x, y \in B_R(p)$. As explained above, the theorem follows from this. □

Exercise 10.6.3. Prove that, if X is a locally compact (but not necessarily complete) space of curvature bounded below, then all precompact open sets in X have the same Hausdorff dimension.

Remark 10.6.4. In fact, the theorem remains valid without the local compactness assumption. This can be proved using results of Section 10.8; see Exercise 10.8.22 there.

Corollary 10.6.5. *Let X be a complete locally compact space of curvature bounded below. Then $\dim_H(U) = \dim_H(X)$ for any open set $U \subset X$.*

Proof. Since X is complete and locally compact, all balls in X are precompact. By Theorem 10.6.1 all these balls have the same Hausdorff dimension. Since X can be covered by a countable collection of balls (e.g., by balls of integer radii centered at a fixed point), $\dim_H(X)$ equals the dimension of these balls; cf. Proposition 1.7.19. □

10.6.2. Gromov–Bishop inequality. It is easy to see from the proof of dimensional homogeneity that the Hausdorff measure of a larger ball $B_R(p)$ can be estimated from above in terms of the measure of a smaller ball $B_r(p)$. In other words, a lower curvature bound imposes an upper bound for the growth rate of volumes of balls. In fact, the volumes of balls in an Alexandrov space grow not faster than the volumes of balls in the comparison space of the respective dimension. This fact is known as the Gromov–Bishop inequality.

Notation. Recall that a *space form* is a simply connected complete space of constant curvature, i.e., a sphere, a Euclidean space, or a hyperbolic space. For an integer $n \geq 2$, we denote the n-dimensional space form of curvature k by M_k^n. As a special convention, we set $M_k^1 = \mathbb{R}$ if $k \leq 0$ and $M_k^1 = \frac{1}{\sqrt{k}} S^1$ (that is, the circle of length $2\pi/\sqrt{k}$) if $k > 0$. We fix an $n \geq 1$ and denote by V_r^k and S_r^k the volume of the r-ball and the $(n-1)$-dimensional area of the r-sphere in M_k^n.

Theorem 10.6.6 (Gromov–Bishop inequality). *Let X be a locally compact Alexandrov space of curvature $\geq k$ and n be a positive integer. Then for every $p \in X$ the ratio*

$$\frac{\mu_n(B_r(p))}{V_r^k}$$

is nonincreasing in r, where $\mu_n(B_r(p))$ is the n-dimensional Hausdorff measure of a ball of radius $r > 0$ centered at p. In other words, if $R \geq r > 0$, then

$$\frac{\mu_n(B_R(p))}{V_R^k} \leq \frac{\mu_n(B_r(p))}{V_r^k}.$$

We give a complete proof only for $k = 0$ (and this proof is essentially the same as that of dimensional homogeneity). In the general case, we give a proof based on the formula

$$\mu_n(B_r(p)) = \int_0^r \mu_{n-1}(S_t(p))\,dt \tag{10.5}$$

where $S_t(p)$ is the sphere of radius t centered at p. If X is a Riemannian manifold, (10.5) is a simple particular case of the so-called coarea formula (Theorem 3.2.11 in [**Fe**]). For general Alexandrov spaces of curvature bounded below formula (10.5) is still valid but its proof requires hard technical work.

An alternative approach is to modify the proof below so that it uses finite sums instead of the integral formula (10.5). Namely, one can split $B_r(p)$ into thin layers by a collection of spheres of appropriate radii and replace the inequalities involving $(n-1)$-dimensional volumes of spheres by similar ones with n-volumes of these layers. Then, replacing the integral in (10.5) by the sum of volumes of the layers, one can obtain the desired inequality with any given precision. This modification is straightforward but the resulting argument is full of irrelevant technical details about choosing the radii, estimating correction terms and so on. So we present a more clear argument with the formula (10.5), leaving a more elementary (but technically more involved) method as exercise to the reader.

We begin however with a simple proof in the case $k = 0$.

Proof of Theorem 10.6.6 for $k = 0$. Let $f\colon X \to X$ be a (r/R)-homothety centered at p (cf. Lemma 10.6.2). Clearly f maps $B_R(p)$ to $B_r(p)$. Lemma 10.6.2 implies that f is injective and its inverse f^{-1} is Lipschitz with Lipschitz constant R/r. Therefore

$$\mu_n(B_R(p)) \le (R/r)^n \mu_n(B_r(p)) = \frac{V_R^0}{V_r^0}\,\mu_n(B_r(p))$$

and the theorem follows. □

Note that for $k \neq 0$ the same argument proves an inequality of the form $\mu_n(B_R(p)) \le C(r,R,k)\mu_n(B_r(p))$ for some (nonoptimal) constant $C(r,R,k)$. This is sufficient for many purposes.

To prove the theorem in full generality, we need the following lemma which is similar to the Gromov–Bishop inequality but involves spheres instead of balls.

Lemma 10.6.7. *Let $S_t(p)$ denote the sphere of radius t centered at $p \in X$. Then*

$$\frac{\mu_{n-1}(S_r(p))}{\mu_{n-1}(S_R(p))} \geq \frac{S_r^k}{S_R^k} \tag{10.6}$$

if $R \geq r > 0$.

Proof. We will prove the inequality for $k = 1$. By means of re-scaling, the statement then follows for any $k > 0$. To obtain a proof for $k < 0$, simply replace sines in the formulas by hyperbolic sines.

Consider an (r/R)-homothety f (see Lemma 10.6.2) restricted to the sphere $S_R(p)$, and let f_0 be the analog of f in the 1-plane. We consider f_0 as a map from the R-sphere in the 1-plane to the r-sphere in the 1-plane. The derivative of f_0 multiplies lengths of all tangent vectors by $\sin r/\sin R$. Hence f_0 is locally Lipschitz with local Lipschitz constant $\sin r/\sin R$ and f_0^{-1} is locally Lipschitz with local Lipschitz constant $\sin R/\sin r$. Therefore for any $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\frac{|f_0(x)f_0(y)|}{|xy|} \geq (1-\varepsilon)\frac{\sin r}{\sin R}$$

whenever $|xy| < \delta$. By triangle comparison the same statement holds for f.

Let us partition $S_R(p)$ into finitely or countably many sets D_i of diameter less than δ. The above inequality implies that

$$\mu_{n-1}(f(D_i)) \geq (1-\varepsilon)\left(\frac{\sin r}{\sin R}\right)^{n-1};$$

hence

$$\begin{aligned}\mu_{n-1}(S_r(p)) \geq \sum \mu_{n-1}(f(D_i)) &\geq (1-\varepsilon)\left(\frac{\sin r}{\sin R}\right)^{n-1} \mu_{n-1}(S_R(p)) \\ &= (1-\varepsilon)\,\frac{S_r^k}{S_R^k}\,\mu_{n-1}(S_R(p)).\end{aligned}$$

Since ε is arbitrary, the lemma follows. □

Proof of Theorem 10.6.6. It is a matter of simple computation to derive the Gromov–Bishop inequality from (10.6) and (10.5). Define $V(t) = \mu_n(B_t(p))$ and $S(t) = \mu_{n-1}(S_t(p))$. Integrating the inequality (10.6) in r over $[0, R]$ yields

$$\frac{V(R)}{S(R)} \geq \frac{V_R^k}{S_R^k},$$

or, equivalently (with R replaced by t),

$$\frac{d}{dt}\log V(t) = \frac{S(t)}{V(t)} \leq \frac{S_t^k}{V_t^k} = \frac{d}{dt}\log V_t^k.$$

Integrating this in t over $[r, R]$ yields the assertion of the theorem.

To justify the above computations with derivatives, observe that (10.6) implies that the function $t \mapsto S(t)$ is a product of the continuous function $t \mapsto S_t^k$ and a monotone function. Hence for all t except at most countably many values, $S(t)$ is continuous at t and equals dV/dt. □

10.6.3. Bishop inequality. Suppose that X is an n-dimensional Riemannian manifold of curvature $\geq k$. Fix a $p \in X$. Then volumes of small balls in X centered at p are almost Euclidean; hence

$$\frac{\mu_n(B_r(p))}{V_r^k} \to 1, \qquad r \to 0.$$

Then the Gromov–Bishop inequality implies that $\mu_n(B_R(p)) \leq V_R^k$ for any $R > 0$. This fact (known as the Bishop inequality) can be generalized to all Alexandrov spaces.

Theorem 10.6.8 (Bishop inequality). *Let n be a positive integer and X an n-dimensional Alexandrov space of curvature $\geq k$. Then for every $p \in X$ and every $r > 0$*

$$\mu_n(B_r(p)) \leq V_r^k,$$

where V_r^k is volume of a ball of radius r in the space form M_k^n (see the notations in Subsection 10.6.2).

If $k > 0$, the Bishop inequality and the upper bound for the diameter (Theorem 10.4.1) imply that the volume of X is no greater than the volume of the n-dimensional sphere of radius $1/\sqrt{k}$. In other words, the following statement holds.

Corollary 10.6.9. *If X is an n-dimensional Alexandrov space of curvature $\geq k$ where $k > 0$, then*

$$\mu_n(X) \leq \frac{\mu_n(S^n)}{k^{n/2}}.$$

Theorem 10.6.8 immediately follows from the next proposition. The proposition says that the ball $B_r(p)$ is "not greater" than the r-ball in M_k^n in the sense that there is a noncontracting map from one to the other.

Proposition 10.6.10. *Let X be an n-dimensional Alexandrov space of curvature $\geq k$, and $p \in X$. Then there exists a map from $f \colon X \to M_k^n$ such that $|f(x)f(y)| \geq |xy|$ for all $x, y \in X$ (i.e., f is noncontracting) and $|f(p)f(x)| = |px|$ for all $x \in X$.*

Proof. We argue by induction in n using the following fact (Theorem 10.8.6) proved later in this chapter: if $n \geq 2$ and X is an n-dimensional Alexandrov

space, then the space of directions $\Sigma_p(X)$ at any point $p \in X$ is an $(n-1)$-dimensional Alexandrov space of curvature ≥ 1; if X is one-dimensional, then $\Sigma_p(X)$ consists of one or two points. This fact and the inductive hypothesis imply that there exists a noncontracting map from $\Sigma_p(X)$ to $M_1^{n-1} = S^{n-1}$.

Denote by $K_p^k(X)$ the k-cone (cf. Subsection 10.2.1) over the space of directions $\Sigma_p(X)$. The following lemma follows immediately from angle comparison.

Lemma 10.6.11. *Let X be an Alexandrov space of curvature $\geq k$ and $p \in X$. Consider the map $\log_p : X \to K_p^k(X)$ which maps p to the origin of the cone and every point $x \in X \setminus \{p\}$ to the pair $(\xi, |px|)$ where ξ is the direction of some (arbitrarily chosen) shortest path $[px]$. This map $\log_p$ is noncontracting.*

By inductive hypothesis, we have a noncontracting map from $\Sigma_p(X)$ to S^{n-1}. It naturally extends to a noncontracting map $F : K_p^k(X) \to M_k^n$ preserving the distance from the origin (note that M_k^n is the k-cone over S^{n-1}). Then the map $F \circ \log_p$ is a desired noncontracting map from X to M_k^n. □

Exercise 10.6.12. Prove that the inequality in Theorem 10.6.8 turns into equality if and only if $B_r(p)$ is isometric to the r-ball in M_n^k. In particular, if the equality takes place for all r, then X is isometric to M_k^n.

Remark 10.6.13. The Gromov–Bishop and Bishop inequalities are also valid for Riemannian manifolds with a lower bound for Ricci curvature.

10.6.4. Appendix: Rough dimension. Here we introduce one more notion which simplifies considerations in many cases. It is especially useful in the case when a space is not supposed to be locally compact and finite-dimensional. Let X be a metric space. For every $\varepsilon > 0$, denote by $\beta_X(\varepsilon)$ the maximal cardinality of an ε-separated subset of X, that is, a collection of points $p_i \in X$ such that $|p_ip_j| \geq \varepsilon$ for all $i \neq j$. The case $\beta_X(\varepsilon) = \infty$ is not excluded.

Definition 10.6.14. The rough d-dimensional volume $\mathrm{Vol}_d X$ is

$$\limsup_{\varepsilon \to 0} \varepsilon^d \beta_X(\varepsilon).$$

Rough dimension $\dim_r$ is defined by the equality

$$\dim_r(X) = \sup\{d\colon \mathrm{Vol}_d X = \infty\}.$$

Exercise 10.6.15. Prove that

1. $\dim_H X \leq \dim_r X$, where $\dim_H$ denotes Hausdorff dimension.

Hint: this is obvious.

2. For any map $f\colon X \to Y$ such that $|f(x)f(y)| \geq c|xy|$ for some constant c, one has $\dim_r f(X) \geq \dim_r X$.

3. $\dim_r(X) = \inf\{d\colon \operatorname{Vol}_d(X) = 0\}$;

4. If $f\colon X \to Y$ is a Lipschitz map, then $\dim_r f(X) \leq \dim_r X$. In particular, $\dim_r f(X) = \dim_r X$ if f is bi-Lipschitz.

10.7. Gromov–Hausdorff Limits

In this section we consider the class of Alexandrov spaces as a subset of the Gromov–Hausdorff "space of metric spaces". First of all, Toponogov's Theorem makes it possible to prove that Alexandrov spaces form a closed set (note that the similar statement about spaces of curvature $\leq k$ is false).

Proposition 10.7.1. *A Gromov–Hausdorff limit of Alexandrov spaces of curvature $\geq k$ is itself a space of curvature $\geq k$. The same is true for Gromov–Hausdorff limits of pointed spaces (in the sense of Definition 8.1.1).*

Proof. By Theorem 7.5.1, a limit of length spaces is a length space. So it suffices to verify the quadruple condition (cf. Proposition 10.1.1) for a limit space. Let $X_n \xrightarrow[GH]{} X$ where $\{X_n\}$ is a sequence of spaces of curvature $\geq k$. Then for every quadruple $(a; b, c, d)$ in X there is a sequence of quadruples $(a_n; b_n, c_n, d_n)$ in X_n such that $|a_n b_n| \to |ab|$, $|a_n c_n| \to |ac|$, and so on. Then the quadruple condition for $(a; b, c, d)$ follows by continuity from that for $(a_n; b_n, c_n, d_n)$. □

The main result of this section is the following Gromov's Compactness Theorem. For every $k \in \mathbb{R}$, $D > 0$ and an integer $n \geq 0$ we define

$$\mathfrak{M}(n,k) = \{X : X \text{ is an AS of curvature} \geq k \text{ and } \dim_H(X) \leq n\},$$
$$\mathfrak{M}(n,k,D) = \{X \in \mathfrak{M}(n,k) : \operatorname{diam}(X) \leq D\}.$$

Theorem 10.7.2 (Gromov's compactness theorem). *The class $\mathfrak{M}(n,k,D)$, regarded with the Gromov–Hausdorff metric, is compact. In particular, $\mathfrak{M}(n,k)$ is compact if $k > 0$.*

An alternative formulation is: $\mathfrak{M}(n,k)$ is boundedly compact, i.e., all its closed bounded subsets are compact.

Some ingredients of the theorem will be proved later. Namely, in Section 10.8 we will show that

- All finite-dimensional Alexandrov spaces are locally compact (Theorem 10.8.1). Therefore all bounded Alexandrov spaces are compact, so $\mathfrak{M}(n,k)$ is indeed a subset of the Gromov–Hausdorff metric space.

- The dimension of a finite-dimensional Alexandrov space is always an integer (Theorem 10.8.2).
- A Gromov–Hausdorff limit of Alexandrov spaces of dimension $\leq n$ has dimension $\leq n$ as well (Corollary 10.8.25). Since the diameter is a continuous function, it follows that $\mathfrak{M}(n,k)$ is a closed set in the Gromov–Hausdorff space.

Given all this, it is sufficient to prove the following

Proposition 10.7.3. *For any n, k, D, the class of all Alexandrov spaces X of curvature $\geq k$ with* $\dim_H(X) = n$ *and* $\operatorname{diam}(X) \leq D$ *is precompact.*

Proof. According to Theorem 7.4.15, it suffices to prove that for any $\varepsilon > 0$ there is an $N = N(\varepsilon, n, k, D) > 0$ such that a space from this class cannot contain an ε-separated set of more than N points.

Let X be an n-dimensional Alexandrov space of curvature $\geq k$ with $\operatorname{diam}(X) \leq D$. By Proposition 10.6.10, there is a noncontracting map from X to the ball of radius D in the space form M_k^n. Let N be the maximal possible cardinality of an ε-separated set in the D-ball in M_k^n. Then X cannot contain an ε-separated set of more than N points, because the image of an ε-separated set under a noncontracting map is ε-separated. □

Remark 10.7.4. There is another proof which appears simpler when all necessary components from the next sections are substituted. Instead of Proposition 10.6.10, one needs the Gromov–Bishop inequality (a simple version with nonsharp constants is sufficient), and the following fact implied by the Gromov–Bishop inequality and Theorem 10.8.3: the volume μ_n of an n-dimensional bounded Alexandrov space is positive and finite.

Now let X be as above. Then by the Gromov–Bishop inequality we have

$$\mu_n(B_\varepsilon(p)) \geq c(n,k,D)\mu_n(B_D(p)) = \mu_n(X)$$

for any $p \in X$ (note that $\overline{B}_D(p) = X$ since $\operatorname{diam}(X) \leq D$). Therefore X cannot contain more than $N = c(n,k,D)^{-1}$ disjoint ε-balls, or, equivalently, cannot contain a (2ε)-separated set of more than N points. The proposition follows.

Remark 10.7.5. The method described in the previous remark also proves the following statement: the class of Riemannian n-manifolds of Ricci curvature $\geq k$ and of diameter no greater than D is precompact in the Gromov–Hausdorff space (because the Gromov–Bishop inequality works with lower bounds for the Ricci curvature). However, no reasonable description is known for the closure of this class in the Gromov–Hausdorff space.

Remark 10.7.6. A limit of n-dimensional spaces may have dimension strictly less than n. For example, for any compact nonnegatively curved

space X rescaled spaces $\{\lambda X\}$ are nonnegatively curved and converge to a point as $\lambda \to 0$. It is possible to make the dimension drop in the limit keeping both lower and upper curvature bounds. For example, "thin" flat tori $S^1 \times (\lambda S^1)$ converge to the circle as $\lambda \to 0$.

Sequences of Alexandrov spaces whose limits have lower dimension are called *collapsing sequences.*

10.8. Local Properties

Now we pass to local properties of finite-dimensional Alexandrov spaces (recall that by dimension we mean the Hausdorff dimension). This section is very technical. For the first reading we recommend skipping long proofs in subsection 10.8.2.

10.8.1. Formulations. Below are some of the formulations that we prove (or explain how to prove) in this and the next sections. The first statement in the list is the local compactness (the property that we assumed in so many places to simplify technical details in proofs).

Theorem 10.8.1. *All finite-dimensional Alexandrov spaces are locally compact.*

Theorem 10.8.2. *The Hausdorff dimension of any Alexandrov space is an integer or infinity.*

Theorem 10.8.3. *An n-dimensional Alexandrov space contains an open dense set which is an n-dimensional manifold. Moreover, there is an open dense set such that every its point has a neighborhood bi-Lipschitz homeomorphic to an open region in $\mathbb{R}^n$.*

These three theorems are proved in the end of Subsection 10.8.2; see Corollaries 10.8.20, 10.8.21 and 10.8.23.

Theorem 10.8.3 is the first step in the study of the local topological structure of an Alexandrov space. According to it, a space is divided into the set of topologically regular points (those having Euclidean neighborhoods), and a nowhere dense set of topological singularities. To obtain more detailed information, one can study the set of singular points; some results about it are discussed later in Section 10.10.

Another way to obtain a classification of points into good and bad ones is based on tangent cones. Namely, a point is said to be *regular* if the tangent cone at it is Euclidean, i.e., isometric to $\mathbb{R}^n$, or, equivalently, the space of directions is isometric to the standard sphere S^{n-1}.

Theorem 10.8.4. *Let X be an n-dimensional Alexandrov space and $p \in X$. Then the following three assertions are equivalent.*

(1) *p is regular.*
(2) *For any $\varepsilon > 0$ there is a neighborhood U of p at a Lipschitz distance less than ε from some open set in $\mathbb{R}^n$. (See Section 7.2 for the definition of Lipschitz distance.)*
(3) *The Gromov–Hausdorff tangent cone of X at p is isometric to $\mathbb{R}^n$. In other words, the dilated balls $\frac{1}{\varepsilon}B_\varepsilon(p)$ converge in the Gromov–Hausdorff sense to the unit ball of $\mathbb{R}^n$ as $\varepsilon \to 0$.*

This theorem is composed of Theorem 10.9.3 and Theorem 10.9.16. In fact, we prove a more general statement: if the space of directions $\Sigma_p(X)$ is close to S^{n-1} in the Gromov–Hausdorff metric, then a neighborhood of p is close to an open set in $\mathbb{R}^n$ in the Lipschitz metric.

Theorem 10.8.5. *Let X be an n-dimensional Alexandrov space. Then the set of regular points is dense in X, and moreover is an intersection of countably many open sets.*

This theorem is proved as Corollary 10.9.13. In fact, the set of nonregular points has Hausdorff dimension no greater than $n - 1$, but we do not prove this. See Section 10.10 for more detailed formulations and discussion.

Spaces of directions and tangent cones are considered in Section 10.9. Among other things, we prove the following theorem (see Corollary 10.9.6).

Theorem 10.8.6. *Let X be an n-dimensional Alexandrov space with $n \geq 2$. Then for any point $p \in X$ the space of directions $\Sigma_p(X)$ is an Alexandrov space of curvature ≥ 1 and of dimension $n - 1$. A space of directions of a one-dimensional Alexandrov space consists of one or two points.*

This theorem sounds very similar to the respective fact (Theorem 9.1.44) about spaces of curvature bounded above. However the proof for the case of curvature bounded below is not that easy. Even the fact that the metric of $\Sigma_p(X)$ is intrinsic appears difficult to prove. We prove Theorem 10.8.6 using Gromov–Hausdorff convergence. The core facts are: $\Sigma_p(X)$ is compact, and the tangent cone $K_p(X)$ is also the Gromov–Hausdorff tangent cone of X at p. Then the desired properties of $\Sigma_p(X)$ follow from the respective properties of $K_p(X)$, which in turn follow from general properties of Gromov–Hausdorff limits.

Remark 10.8.7. A common technique when using a space of directions is to carry out an induction on dimension. This is the reason why we do not take our usual approach "prove only for locally compact spaces" in this section. Even if we did, we could not avoid dealing with possibly not locally compact spaces when passing down to the space of directions.

Note that compactness of the space of directions does not follow from (local) compactness of a space itself, if one does not assume that the dimension is finite. For example, consider a direct metric product of infinitely many spaces X_i where X_i is a sphere of radius $1/i$ and i ranges over all positive integers. This product is a compact length space of nonnegative curvature; however its space of direction at any point is not compact (it contains an infinite $(\pi/2)$-separated set and in fact is isometric to a sphere in an infinite-dimensional Hilbert space).

Exercise 10.8.8. Give a definition of a direct metric product of a (good enough) countable collection of metric spaces, and prove the statements made in this remark.

We fix a number k and let X denote an Alexandrov space of curvature $\geq k$. Since the formulations do not depend on k, we assume that $k \leq 0$ (to make sure that comparison triangles always exist). Since our considerations are local, we may assume k is arbitrarily close to 0 by means of re-scaling. For simplicity we assume that $k = 0$. Making the proofs work for a negative k is a matter of introducing negligibly small correction terms in the formulas.

10.8.2. Strainers. The notion of a strainer is a useful technical tool which somewhat resembles the notion of an orthogonal frame in $\mathbb{R}^n$, and is used for the same purpose—to introduce a (local) coordinate system. To see how one may come up with a definition of a strainer, consider a point $p \in X$ such that $\Sigma_p(X)$ is isometric to S^{n-1}. The sphere contains n pairs of points such that the (angular) distance between points in each pair equals π, and the distances between points from different pairs all equal $\pi/2$. These points correspond to n curves $\gamma_1, \dots, \gamma_n$ in X passing through p and having mutually orthogonal directions at p, similarly to the coordinate axes in $\mathbb{R}^n$.

Then one can introduce coordinates in a neighborhood of p as follows: the coordinates of a point q equal the distance $|pq|$ multiplied by the cosines of the angles between a shortest path $[pq]$ and the "coordinate axes". This gives the standard coordinates if X is the Euclidean space, but in the general case these "coordinates" may even fail to be continuous (e.g., if a shortest path $[pq]$ is not unique). It is clear how to get rid of discontinuity problems once and forever: fix a point a_i in each γ_i and use comparison angles ($\tilde{\measuredangle} qpa_i$) instead of angles. (In fact, this does not solve all the problems, only some of them. The coordinates that we actually introduce later are slightly different.) After this, "coordinate axes" and even their directions are no longer needed, and one formulates the following

Definition 10.8.9. A point $p \in X$ is an (m, ε)*-strained point* if there are m pairs of points (a_i, b_i) in X such that

$$\begin{aligned} \tilde{\angle} a_i p b_i &> \pi - \varepsilon, \\ \tilde{\angle} a_i p a_j &> \tfrac{\pi}{2} - 10\varepsilon, \\ \tilde{\angle} a_i p b_j &> \tfrac{\pi}{2} - 10\varepsilon, \\ \tilde{\angle} b_i p b_j &> \tfrac{\pi}{2} - 10\varepsilon \end{aligned}$$

for all $i, j \in \{1, \ldots, m\}$, $i \neq j$. The collection $\{(a_i, b_i)\}$ itself is called an (m, ε)*-strainer* for p.

The constant 10 in the definition is not very important. It is introduced only to simplify the formulation of one of the subsequent lemmas.

Remark 10.8.10. Recall that $\tilde{\angle} a_i p b_i + \tilde{\angle} a_i p a_j + \tilde{\angle} a_j p b_i \leq 2\pi$ by the quadruple condition. It follows that the angles $\tilde{\angle} a_i p a_j$ and $\tilde{\angle} a_j p b_i$ (and, similarly, $\tilde{\angle} b_i p b_j$) are not only greater than $\frac{\pi}{2} - 10\varepsilon$, they are also less than $\frac{\pi}{2} + 11\varepsilon$; i.e., they are actually close to $\frac{\pi}{2}$.

One does not have to send ε to zero. For our purposes it is sufficient to fix once and forever a small ε, say, $\varepsilon < \varepsilon_0 = \frac{1}{100m}$. If ε satisfies this inequality, we omit it and write simply "*m-strainer*" and "*m-strained point*". Obviously the set of (m, ε)-strained points is open for any fixed m and ε.

Definition 10.8.11. Let X be an Alexandrov space. The *strainer number* of X is the supremum of numbers m such that there exists an m-strainer in X.

A *strainer number* at a point $x \in X$ is the supremum of numbers m such that any neighborhood of x contains an m-strained point.

We will show in this section that the strainer number equals the Hausdorff dimension of the space (in particular, the latter is an integer or infinity). Right now, observe that X admits an $(1, \varepsilon)$-strainer for any $\varepsilon > 0$ unless X is a single point. To prove this, pick any two points a, b and let p be an almost midpoint (more precisely, a δ-midpoint with δ depending on $|ab|$ and ε); then (a, b) is the desired $(1, \varepsilon)$-strainer for p.

The following proposition tells us that the notion of a strained point is local, and moreover can be formulated in terms of the space of directions.

Proposition 10.8.12. *1. If $\{(a_i, b_i)\}$ is an (m, ε)-strainer for $p \in X$, and $a_i' \in [a_i p]$, $b_i' \in [b_i p]$ for $i = 1, \ldots, m$, then $\{(a_i', b_i')\}$ is an (m, ε)-strainer for p as well. Even if shortest paths do not exist, the same is true for points a_i' and b_i' taken in suitable almost shortest paths. In particular, there exists an (m, ε)-strainer $\{(a_i', b_i')\}$ with arbitrarily small distances $|pa_i'|$ and $|pb_i'|$.*

2. If $\{(a_i, b_i)\}$ is an (m, ε)-strainer for p and shortest paths $[pa_i]$ and $[pb_i]$ exist for all i, then the angles between these shortest paths satisfy the inequalities

$$\begin{aligned} \measuredangle a_i p b_i &> \pi - \varepsilon, \\ \measuredangle a_i p a_j &> \tfrac{\pi}{2} - 10\varepsilon, \\ \measuredangle a_i p b_j &> \tfrac{\pi}{2} - 10\varepsilon, \\ \measuredangle b_i p b_j &> \tfrac{\pi}{2} - 10\varepsilon \end{aligned}$$

for all $i, j \in \{1, \ldots, m\}$, $i \neq j$.

3. Conversely, if the inequalities from the second statement hold for some points p, $\{a_i\}$ and $\{b_i\}$ ($i = 1, \ldots, m$), then p is an (m, ε)-strained point.

Proof. The first statement follows from the monotonicity of angles. The second one is obtained as the limit of the first one as a_i' and b_i' converge to p along the respective shortest paths. To prove the third statement, observe that $\{(a_i', b_i')\}$ is an (m, ε)-strainer if a_i' and b_i' are sufficiently close to p in the respective geodesics $[pa_i]$ and $[pb_i]$. □

The following two lemmas are simple but important technical facts that are used everywhere in this section. The first one is an analog of the fact that the sum of adjacent angles equals π, but with a strainer and comparison angles instead of a geodesic and angles between directions.

Lemma 10.8.13. *Let (a, b) be a $(1, \varepsilon)$-strainer for p, and $q \in X$ be such that*

$$|pq| < \frac{\varepsilon}{4} \min\{|pa|, |pb|\}.$$

Then

$$|\tilde{\measuredangle} apq + \tilde{\measuredangle} bpq - \pi| < \varepsilon.$$

In particular, if shortest paths $[pa]$, $[pb]$ and $[pq]$ exist, the angles between them are close to the comparison angles:

$$\begin{aligned} 0 &< \measuredangle apq - \tilde{\measuredangle} apq < 2\varepsilon, \\ 0 &< \measuredangle bpq - \tilde{\measuredangle} bpq < 2\varepsilon. \end{aligned}$$

Proof. By the quadruple condition for $(p; a, b, q)$ we have

$$\tilde{\measuredangle} apq + \tilde{\measuredangle} bpq \leq 2\pi - \tilde{\measuredangle} apb < \pi + \varepsilon.$$

It remains to prove that $\tilde{\measuredangle} apq + \tilde{\measuredangle} bpq > \pi - \varepsilon$. Suppose the contrary, and place the comparison triangles $\triangle\bar{a}\bar{p}\bar{q}$ and $\triangle\bar{b}\bar{p}\bar{q}$ in different half-planes with respect to $[\bar{p}\bar{q}]$. Then

$$\measuredangle \bar{a}\bar{p}\bar{b} = \measuredangle \bar{a}\bar{p}\bar{q} + \measuredangle \bar{b}\bar{p}\bar{q} < \pi - \varepsilon < \tilde{\measuredangle} apb;$$

hence $|ab| > |\bar{a}\bar{b}|$. The inequality $|pq| < \frac{\varepsilon}{4}|pa|$ implies $\tilde{\angle}paq < \arcsin(\frac{\varepsilon}{4}) < \frac{\varepsilon}{2}$; then

$$\angle\bar{a}\bar{q}\bar{p} = \pi - \tilde{\angle}paq - \tilde{\angle}apq > \pi - \frac{\varepsilon}{2} - \tilde{\angle}apq,$$

and similarly $\angle\bar{b}\bar{q}\bar{p} > \pi - 2\delta - \tilde{\angle}bpq$. Thus

$$\angle\bar{a}\bar{q}\bar{p} + \angle\bar{b}\bar{q}\bar{p} > 2\pi - \varepsilon - (\tilde{\angle}apq + \tilde{\angle}bpq) > \pi$$

(i.e., the angle at $\bar{q}$ of the quadrangle $\bar{a}\bar{p}\bar{b}\bar{q}$ is greater than π). Since $|ab| > |\bar{a}\bar{b}|$, it follows that

$$\tilde{\angle}aqb > \angle\bar{a}\bar{q}\bar{b} = 2\pi - \angle\bar{a}\bar{q}\bar{p} - \angle\bar{b}\bar{q}\bar{p},$$

or, equivalently, $\tilde{\angle}aqb + \tilde{\angle}aqp + \tilde{\angle}bqp > 2\pi$, contrary to the quadruple condition for $(q; a, b, p)$.

The second statement of the lemma (about angles between shortest paths) follows from the first one. Indeed, the quadruple condition implies that $\angle apq + \angle bpq \le 2\pi - \angle apb < \pi + \varepsilon$. Therefore

$$(\angle apq - \tilde{\angle}apq) + (\angle bpq - \tilde{\angle}bpq) < (\pi + \varepsilon) - (\pi - \varepsilon) = 2\varepsilon.$$

Since both terms $\angle apq - \tilde{\angle}apq$ and $\angle bpq - \tilde{\angle}bpq$ are positive, it follows that they are bounded above by 2ε. □

Lemma 10.8.14. *Let $p, a_1, a_2, b_1, b_2 \in X$ be such that (a_1, b_1) and (a_2, b_2) are $(1, \varepsilon)$-strainers for p,*

$$|a_2b_2| < \frac{\varepsilon}{4}\min\{|pa_1|, |pb_1|\}$$

and

$$\big||a_1a_2| - |a_1b_2|\big| < \varepsilon|a_2b_2|.$$

Then $\{(a_1, b_1), (a_2, b_2)\}$ is a $(2, \varepsilon)$-strainer for p.

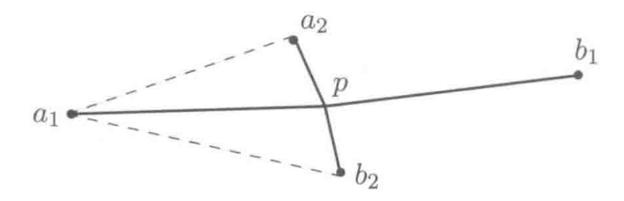


Figure 10.2: Lemma 10.8.14.

Proof. It is sufficient to prove that the comparison angles $\tilde{\angle}a_1pa_2$ and $\tilde{\angle}a_1pb_2$ are greater than $\frac{\pi}{2} - 8\varepsilon$. Indeed, if this is true, then the quadruple condition for $(p; a_1, a_2, b_2)$ implies that these two angles are less than $\frac{\pi}{2} + 9\varepsilon$ (compare with Remark 10.8.10 after the definition of a strainer). Then

Lemma 10.8.13 implies that the angles $\tilde{\angle} a_2pb_1$ and $\tilde{\angle} b_1pb_2$ are greater than $\frac{\pi}{2} - 10\varepsilon$.

Suppose the contrary, for example; let $\tilde{\angle} a_1pa_2 < \frac{\pi}{2} - 8\varepsilon$. Then

$$\tilde{\angle} a_2pb_1 > \pi - \varepsilon - \tilde{\angle} a_1pa_2 > \frac{\pi}{2} + 7\varepsilon$$

by Lemma 10.8.13,

$$\tilde{\angle} b_1pb_2 < 2\pi - \tilde{\angle} a_2pb_2 - \tilde{\angle} b_1pb_2 < \frac{\pi}{2} - 6\varepsilon$$

by the quadruple condition for $(p; a_2, b_1, b_2)$, and

$$\tilde{\angle} a_1pb_2 > \pi - \varepsilon - \tilde{\angle} b_1pb_2 > \frac{\pi}{2} + 5\varepsilon$$

by Lemma 10.8.13. We are going to show that the difference $|a_1b_2| - |a_1a_2|$ is too large to satisfy the last condition of the lemma.

Here and later in this section we need the following elementary fact: if $\triangle xyz$ is a triangle in $\mathbb{R}^2$, then

$$|xz| > |xy| + |yz| \cdot \sin(\angle xyz - \tfrac{\pi}{2}). \tag{10.7}$$

To prove this inequality, observe that the right-hand side of the inequality equals the distance between x and the projection of z in the line xy.

Since $|pa_2| < |a_2b_2| < \frac{\varepsilon}{4}|pa_1|$, we have $\tilde{\angle} pa_1a_2 < \frac{\varepsilon}{2}$; therefore

$$\tilde{\angle} a_1a_2p > \pi - \tilde{\angle} a_1pa_2 - 2\delta > \frac{\pi}{2} + 7\varepsilon.$$

Applying (10.7) to the comparison triangle for $\triangle a_1a_2p$, we obtain that $|a_1p| > |a_1a_2| + \sin(7\varepsilon)|a_2p|$. Similarly the inequality $\tilde{\angle} a_1pb_2 > \frac{\pi}{2} + 5\varepsilon$ implies that $|a_1b_2| > |a_1p| + \sin(5\varepsilon)|pb_2|$. Thus

$$|a_1b_2| > |a_1a_2| + \sin(5\varepsilon)(|a_2p| + |p_2|) > |a_1a_2| + \varepsilon|a_2b_2|,$$

which contradicts the last condition of the lemma. □

Now we introduce local coordinates in a neighborhood of a strained point. We use distances to the points a_i as coordinate functions. More formally, if $\{(a_i, b_i)\}$ is an m-strainer for p, we define the map $f : U \to \mathbb{R}^m$, where U is a neighborhood of p, by

$$f(x) = (|xa_1|, \dots, |xa_m|).$$

The functions $x \mapsto |xa_i|$ (and the map f itself) are referred to as *distance coordinates* associated with the strainer $\{(a_i, b_i)\}$. Obviously f is a Lipschitz map because its coordinate functions are.

The neighborhood U is supposed to be sufficiently small. First, the set of $x \in X$ such that $\{(a_i, b_i)\}$ is an m-strainer for x is open (recall that an m-strainer is an (m, ε)-strainer with $\varepsilon = \frac{1}{100m}$). So we may assume that U

is contained in this set. Second, we want the diameter of U be significantly smaller than $\min\{|pa_i|, |pb_i|\}$.

If $X = \mathbb{R}^n$ and $m = n$, the level sets of the functions $x \mapsto |xa_i|$ are spheres intersecting almost orthogonally near p. Therefore, for each collection of radii $(r_1, \ldots, r_n)$ near $(|px_1|, \ldots, |px_n|)$ the respective spheres have a unique intersection point near p. The same behavior occurs for a maximal strainer in any Alexandrov space. This statement consist of two parts: that f is an open map (i.e., the image of any open set is open), and that f is an injective map whose inverse map is Lipschitz. We begin with the first part (which holds for any strainer, not only a maximal one).

Proposition 10.8.15. *Let $p \in X$ be an (m, ε)-strained point with $\varepsilon = \frac{1}{100m}$. Then there is a neighborhood U of p such that the distance coordinates form an open map from U to $\mathbb{R}^m$.*

Proof. Let $\{(a_i, b_i)\}$ be an (m, ε)-strainer for p and $f : U \to \mathbb{R}^m$ be the associated distance coordinates. We may assume that $|qa_i| > 1$ and $|qb_i| > 1$ for all $q \in U$, $i = 1, \ldots, m$. This can be achieved by means of re-scaling and choosing U sufficiently small. Furthermore, we assume that U is so small that $\{(a_i, b_i)\}$ is an (m, ε)-strainer for any point $q \in U$. Let us show that $f(U)$ contains a neighborhood of $f(p)$ in $\mathbb{R}^m$. Then the same arguments applied to an arbitrary $q \in U$ instead of p will prove that $f(U)$ contains a neighborhood of q and therefore f is an open map.

Let $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$ be a point in a small neighborhood of $f(p)$. We have to find a point $x \in U$ such that $f(x) = y$. Such a point is obtained by means of consecutive approximations. The idea is the following. Take $x = p$ as the initial approximation. Then, at each step, pick an $i_0 \in \{1, \ldots, m\}$ and move x along $[xa_{i_0}]$ or $[xb_{i_0}]$ until the distance $|xa_{i_0}|$ takes the desired value (i.e., becomes equal y_{i_0}). Since we move x almost orthogonally to the other paths $[xa_i]$, the distances $|xa_i|$ $(i \neq i_0)$ remain almost unchanged. Therefore the point $f(x)$ gets closer to y. On the other hand, the distance between the old and new positions of our point x approximately equals the change in the i_0th coordinate of $f(x)$ or, which is almost the same, the distance between the old and new position of $f(x)$. Since $f(x)$ approaches y quite rapidly, the path that it goes along cannot be long; therefore x also does not run away and the process converges. The formal details follow.

Set $x_0 = p$ and consider a sequence $x_0, x_1, \ldots$ of points where x_{n+1} is chosen recursively depending on x_n as follows. Define $\delta_i = \delta_{i,n} = \big|y_i - |x_n a_i|\big|$ for $i = 1, \ldots, m$ and $\Delta_n = \|y - f(x_n)\|_1 = \sum_{i=1}^m |\delta_{i,n}|$. It will follow by induction that $\Delta_n < \Delta_0$, and we may assume that $\Delta_0 < \varepsilon/10$. Suppose that $x_n \in U$ (this also has to be verified later by induction). Define $\delta = \max_i |\delta_i|$

and choose i_0 for which $|\delta_{i_0}| = \delta$. Clearly $\frac{1}{m}\Delta_n \le \delta \le \Delta_n$. Now let x_{n+1} be a point in the union of shortest paths $[x_n a_{i_0}] \cup [x_n b_{i_0}]$ such that $|x_{n+1}a_{i_0}| = y_{i_0}$; i.e., the i_0-th coordinate of $f(x_{n+1})$ takes the desired value. (If shortest paths do not exist, use "almost shortest" paths instead.) We are going to show that other coordinates of $f(x_{n+1})$ are only slightly different from those of $f(x_n)$.

First, observe that $|x_n x_{n+1}| < 2\delta$. If $|x_n a_{i_0}| < |x_{n+1}a_{i_0}|$, then $x_{n+1} \in [x_n a_{i_0}]$ and we have $|x_n x_{n+1}| = |x_n a_{i_0}| - |x_{n+1}a_{i_0}| = \delta$. Otherwise $x_{n+1} \in [x_n b_{i_0}]$; then $\tilde{\angle} a_{i_0} x_n x_{n+1} \ge \tilde{\angle} a_{i_0} x_n b_{i_0} > \pi - \varepsilon$ by monotonicity of angles, and the inequality $|x_n x_{n+1}| < 2\delta$ follows from (10.7) and the relation $|x_{n+1}a_{i_0}| = |x_n a_{i_0}| + \delta$.

Recall that $\{(a_i, b_i)\}$ is an (m,ε)-strainer for x_n as long as $x_n \in U$, and by Proposition 10.8.12 it remains an (m,ε)-strainer if one replaces a_{i_0} or b_{i_0} by x_{n+1} (depending on whether x_n belongs to $[x_n a_{i_0}]$ or $[x_n b_{i_0}]$). Hence $|\tilde{\angle} a_i x_n x_{n+1} - \frac{\pi}{2}| < 11\varepsilon$ for any $i \ne i_0$ (cf. Remark 10.8.10). Since $|a_i x_n| > 1$ and $|x_n x_{n+1}| < 2\delta$, we have $\tilde{\angle} a_i x_n x_{n+1} < 4\delta < \varepsilon$; hence $|\tilde{\angle} a_i x_{n+1} x_n - \frac{\pi}{2}| < 12\varepsilon$ if $i \ne i_0$. Then (10.7) implies that

$$\big||a_i x_n| - |a_i x_{n+1}|\big| < |x_n x_{n+1}| \cdot \sin(12\varepsilon) < 24\delta\varepsilon$$

and therefore $|\delta_{i,n+1}| < |\delta_{i,n}| + 24\varepsilon\delta$ for all $i \ne i_0$. Thus

$$\Delta_{n+1} < \Delta_n - \delta + 24(m-1)\varepsilon\delta < \Delta_n - \tfrac{1}{2}\delta < (1 - \tfrac{1}{2m})\Delta_n$$

(here we use that $\varepsilon \le \frac{1}{100m}$). Note that we have verified one of our inductive assumptions: $\Delta_{n+1} < \Delta_n < \Delta_0$. Moreover, $\Delta_n < (1-\frac{1}{2m})^n \Delta_0$, so $\{\Delta_n\}$ is a vanishing sequence and $f(x_n) \to y$ as $n \to \infty$, provided that $x_n \in U$ for all n.

It remains to prove that the sequence $\{x_n\}$ does not leave U and is a Cauchy sequence. Recall that

$$|x_n x_{n+1}| < 2\Delta_n < 2\Delta_0(1 - \tfrac{1}{2m})^n$$

while $x_n \in U$; therefore

$$|p x_{n+1}| = |x_0 x_{n+1}| \le \sum_{j=0}^{n} |x_j x_{j+1}| < \Delta_0 \cdot C(m)$$

where $C(m) = \sum_{j=0}^{\infty}(1-\frac{1}{2m})^j < \infty$. We may choose our target $y \in \mathbb{R}^m$ so close to $f(p)$ that $\Delta_0 = \|y - f(p)\|_1 < \operatorname{diam}(U)/C(m)$; then $|p x_{n+1}| < \operatorname{diam}(U)$, and by induction the whole sequence $\{x_n\}$ is contained in U. Since $\sum_{n=1}^{\infty} |x_n x_{n+1}| < \operatorname{diam}(U) < \infty$, the sequence $\{x_n\}$ is Cauchy and its limit belongs in U. This limit is the desired point x such that $f(x) = y$. □

Corollary 10.8.16. *The strainer number of an Alexandrov space is not greater than its Hausdorff dimension.*

Proof. If there is an m-strainer in X, then by the theorem there exists a Lipschitz map from a subset of X onto an open set $V \subset \mathbb{R}^m$. Then $\dim_H(X) \geq \dim_H(V) = m$ because Lipschitz maps do not increase the Hausdorff dimension (Proposition 1.7.19). □

Recall that an m-strainer is defined as an (m, ε)-strainer where ε is an explicit small number, namely, $\varepsilon = \frac{1}{100m}$. It might sound more natural if we required existence of (m, ε)-strainers with arbitrarily small ε. Fortunately, this stronger requirement follows automatically: once any m-strainer is found, its "quality" can be improved to whatever value is needed.

Proposition 10.8.17 (improving a strainer). *If p is an (m, ε)-strained point with $\varepsilon \leq \frac{1}{100m}$, then for any $\varepsilon' > 0$ any neighborhood of p contains an (m, ε')-strained point.*

Proof. Let $\{(a_i, b_i)\}$ be an (m, ε)-strainer for p. Let us move these points (including p) so as to obtain an (m, ε')-strainer. We will abuse notation and denote the new points by the same letters p, a_i, b_i. The improvement of the strainer is done in two stages. First, the angles $\widetilde{\measuredangle} a_i p b_i$ are made very close to π. Then other angles are made very close to $\frac{\pi}{2}$.

To make an angle $\widetilde{\measuredangle} a_{i_0} p b_{i_0}$ close to π, do the following. Consider the distance coordinates $f : U \to \mathbb{R}^m$ associated with $\{(a_i, b_i)\}$ where U is a very small neighborhood of p. Since $f(U)$ is an open set in $\mathbb{R}^m$, one can find two points $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_m)$ in $f(U)$ such that $x_i = y_i$ for all $i \neq i_0$, and $x_{i_0} \neq y_{i_0}$. Choose a and b in U such that $x = f(a)$ and $y = f(b)$; then $|a_i a| = |a_i a|$ for all $i \neq i_0$. Replace a_{i_0} by a, b_{i_0} by b, and p by a midpoint p' between a and b (or an almost midpoint p' such that $\widetilde{\measuredangle} a p' b < \pi - \varepsilon'$). Lemma 10.8.14 applied to points p', a_i, b_i, a, b implies that the new collection of points is an (m, ε)-strainer for the new point $p = p'$.

Applying this construction for $i_0 = 1, \dots, m$ in turn, one obtains an (m, ε)-strainer in which $\widetilde{\measuredangle} a_i p b_i > \pi - \varepsilon'$. (One may wonder why the relation, say, $\widetilde{\measuredangle} a_1 p b_1 > \pi - \varepsilon'$, does not break down while we fix other angles. The answer is easy: the subsequent neighborhoods in which we choose points are much smaller than the first one.)

Now apply the same procedure one more time. Since all (a_i, b_i) are now $(1, \varepsilon')$-strainers, Lemma 10.8.14 implies that the resulting collection of points is an (m, ε')-strainer. □

Theorem 10.8.18. *If $p \in X$ is an m-strained point and m equals the local strainer number at p, then p has a neighborhood which is bi-Lipschitz homeomorphic to an open region in $\mathbb{R}^n$.*

A bi-Lipschitz homeomorphism is provided by distance coordinates associated with any m-strainer. Moreover the Lipschitz constants of this map and its inverse are not greater than $500m$.

Proof. Let $\{(a_i, b_i)\}$ be an m-strainer for a $p \in X$. Assume that there are no $(m+1)$-strained points in a small neighborhood U of p, and let $f : U \to \mathbb{R}^m$ be the map composed from the distance coordinates. We already know that f is an open map (Proposition 10.8.15).

To demonstrate the idea of the proof, let us first show that f is injective (and therefore is a local homeomorphism). Suppose it is not; i.e., there are points $a, b \in U$ such that $|a_i a| = |a_i b|$ for $i = 1, \dots, m$. Then Lemma 10.8.14 implies that an almost midpoint between a and b is $(m+1)$-strained by the original points $\{a_i\}$ and $\{b_i\}$ and the pair (a, b) taken as (a_{m+1}, b_{m+1}).

To prove that f is bi-Lipschitz, we have to turn this argument to an effective estimate for $|f(a)f(b)|$ from below in terms of $|ab|$. By Lemma 10.8.14, for at least one i we have $\big||a_i a| - |a_i b|\big| \geq \frac{\varepsilon}{4}|ab|$ where $\varepsilon = \frac{1}{100m}$; otherwise we have an $(m+1)$-strainer for an almost midpoint between a and b. Then $|f(a)f(b)| \geq \big||a_i a| - |a_i b|\big| \geq \frac{1}{500m}|ab|$; hence the dilatation of f^{-1} is not greater than $500m$.

The dilatation of f itself is not greater than m because the coordinate functions are nonexpanding. □

Remark 10.8.19. The Lipschitz constant $100m$ in the theorem is a poor achievement. A more delicate argument shows that the dilatations of f and f^{-1} are (arbitrarily) close to 1 provided that p is an (m, ε)-strained point with a sufficiently small ε.

Now we are in position to prove some of the theorems formulated in the beginning of this section.

Corollary 10.8.20. *All finite-dimensional Alexandrov spaces are locally compact.*

Proof. Let X be an Alexandrov space and m be its strainer number. We know that $m \leq \dim_H(X) < \infty$. Let $p \in X$ be an m-strained point; then by Theorem 10.8.18 some neighborhood U of X is homeomorphic to $\mathbb{R}^m$. In particular, U is locally compact, so there is an $r > 0$ such that the ball $B_r(p)$ is precompact.

Let us prove that for any $R > 0$ the ball $B_R(p)$ is precompact. Suppose the contrary; then there is an $\varepsilon > 0$ such that $B_R(p)$ contains an infinite ε-separated set S. Recall the construction from Lemma 10.6.2: if the metric is strictly intrinsic, one defines a homothety map $f : B_R(p) \to B_r(p)$ such that $|f(x)f(y)| \geq \frac{r}{R}|xy|$. Namely, the image $f(x)$ of an $x \in B_R(p)$ is a

point in a shortest path $[px]$ such that $|pf(x)| = \frac{r}{R}|px|$. The metric of X is not strictly intrinsic a priori, but one can modify this construction using "almost shortest" paths instead of shortest ones. This way an inequality $|f(x)f(y)| > \frac{r}{2R}|xy|$ can be ensured for points x, y ranging over any given finite set, in particular, any finite subset $S' \subset S$. Thus the ball $B_r(p)$ contains an $\frac{r}{2R}$-separated net $f(S')$ of an arbitrarily large number of points and therefore it is not precompact. This contradiction shows that any ball in X is precompact; hence X is locally compact. □

Corollary 10.8.21. *The Hausdorff dimension of a locally compact Alexandrov space (in particular, of any finite-dimensional one) equals its strainer number.*

Therefore the Hausdorff dimension of any Alexandrov space is an integer or infinity.

Proof. Let X be an Alexandrov space and m be its strainer number. We already know (Corollary 10.8.16) that $m \le \dim_H(X)$. If $m = \infty$, there is nothing else to prove. Otherwise consider a neighborhood U of an m-strained point which is bi-Lipschitz homeomorphic to $\mathbb{R}^m$. Since a bi-Lipschitz homeomorphism preserves the Hausdorff dimension, we have $\dim_H(U) = m$. By dimensional homogeneity (Theorem 10.6.1) it then follows that $\dim_H(X) = \dim_H(U) = m$. □

Exercise 10.8.22. Prove the statement of the above corollary without the local compactness assumption.

Hint: The only case not covered yet is the following: $\dim_H(X) = \infty$ but $m < \infty$. Since $m < \infty$, there is at least one precompact open set (namely, a neighborhood of an m-strained point). Show that the proof of dimensional homogeneity (namely the construction of a homothety) works in this case.

Corollary 10.8.23. *Let X be an n-dimensional Alexandrov space. Then the set of n-strained points is an open dense set in X. Furthermore, this set is an n-dimensional manifold.*

Proof. The fact that the set of n-strained points is an n-manifold follows immediately from Theorem 10.8.18. The first statement of the corollary is equivalent to saying that the local strainer number at any point $p \in X$ equals n. Let m be the local strainer number at p. Then some neighborhood of U of p is bi-Lipschitz homeomorphic to $\mathbb{R}^m$ and hence $\dim_H(U) = m$. Then $m = \dim_H(X) = n$ by dimensional homogeneity. □

Combining Corollary 10.8.23 and Proposition 10.8.17 about improving a strainer, we obtain the following

Corollary 10.8.24. *Let X be an n-dimensional Alexandrov space. Then for any $\varepsilon > 0$ the set of (n, ε)-strained points is an open dense set in X.*

Finally, we prove the estimate for the dimension of a Gromov–Hausdorff limit which was used in the proof of Compactness Theorem 10.7.2 and has numerous other applications.

Corollary 10.8.25. *A Gromov–Hausdorff limit of compact Alexandrov spaces of curvature $\geq k$ and of dimension no greater than n is also an Alexandrov space of curvature $\geq k$ and dimension no greater than n.*

The same is true for Gromov–Hausdorff limits of pointed spaces.

Proof. The proofs for compact and pointed spaces are similar. Let $\{X_i\}$ be a sequence of compact Alexandrov spaces of curvature $\geq k$, $\dim_H(X_i) \leq n$ for all i, $X_i \xrightarrow[GH]{} X$ as $i \to \infty$. By Proposition 10.7.1, X is a (compact) space of curvature $\geq k$. Suppose that $\dim_H(X) > n$. Since $\dim_H(X)$ equals the strainer number of X (Corollary 10.8.21), there is an $(n+1)$-strained point $p \in X$. Fix an $(n+1)$-strainer $\{(a_j, b_j)\}_{j=1}^{n+1}$ for p. As soon as X_i gets sufficiently close to X, one can find points p', a'_j, b'_j whose distances from one another are almost equal to the respective distances between points p, a_j, b_j in X. Then the comparison angles $(\tilde{\angle})$ involving these points in X_i are almost equal to the respective comparison angles in X, because the angles $\tilde{\angle}$ are continuous functions of distances. In particular, the inequalities from the definition of a strained point are satisfied for p', a'_j, b'_j. Thus X_i contains an $(n+1)$-strainer, and therefore $\dim_H(X_i) \geq n+1$, for all large enough i. □

10.9. Spaces of Directions and Tangent Cones

10.9.1. Basic properties.

Proposition 10.9.1. *Let X be a finite-dimensional Alexandrov space. Then for any $p \in X$ the space of directions $\Sigma_p(X)$ is compact.*

Proof. The proof is based on the following

Lemma 10.9.2. *There is a constant $C = C(n, k)$ satisfying the following. If X is an n-dimensional Alexandrov space of curvature $\geq k$, $p \in X$, $0 < r < 1$ and $0 < \varepsilon < 1$, then the ball $B_r(p)$ cannot contain an (εr)-separated set consisting of more than C/ε^n points.*

Proof of the lemma. This lemma is nothing but a refinement of Corollary 10.8.20 about local compactness. All we need is to plug in explicit estimates for the number of points in separated sets.

The statement is trivial for balls in $\mathbb{R}^n$. It then follows for neighborhoods of an n-strained point p, with a constant C written in terms of Lipschitz

constants given in Theorem 10.8.18. Then, with a homothety map from Lemma 10.6.2, the statement is extended to all balls. (One needs, however, an upper bound for the radius to make this work for $k < 0$.) The details follow.

Let p be an n-strained point and r be so small that $U = B_r(p)$ admits a bi-Lipschitz map $f : U \to \mathbb{R}^n$ with $\operatorname{dil} f \leq c_1$ and $\operatorname{dil} f^{-1} \leq c_1$ where $c_1 = 500n$ (cf. Theorem 10.8.18). Then any (εr)-separated set in $B_r(p)$ corresponds to an $(\varepsilon r/c_1)$-separated set in a ball of radius $c_1 r$ in $\mathbb{R}^n$. Such a set cannot contain more than C/ε^n points where $C = c_2 \cdot c_1^2$, c_2 is the respective constant for $\mathbb{R}^n$.

Almost the same constant C works for any ball centered at any point. Indeed, consider an (εr)-separated set $S \subset B_r(p)$ where $p \in X$ is an arbitrary point and $0 < r < 1$. Find an n-strained point q near p and a small ball $U = B_\delta(q)$ satisfying Theorem 10.8.18. Consider a $(\delta/2r)$-homothety $f : B_{2r}(q) \to B_\delta(q)$. We have $|f(x)f(y)| \geq (\delta/2r) \cdot |xy|$ for all x, y; hence $f(S)$ is an $(\varepsilon\delta/2)$-separated set in $B_\delta(q)$. Therefore the cardinality of $f(S)$, and that of S, are no greater than $C(\varepsilon/2)^n$.

These formulas should be slightly modified if $k < 0$. It suffices to multiply the constant by $\sinh^m(-k)$ which corresponds to the correction term $\sinh(-k)$ arising in the lower bound for $|f(x)f(y)|$ in the unit ball of the k-plane. □

Now continue the proof of the proposition. We will show that for any $\varepsilon > 0$ the space of directions cannot contain an ε-separated set of more than C/ε^n points where C is some constant (almost the same one as in the lemma). Let $S = \{x_i\}_{i=1}^N$ be an ε-separated set in $\Sigma_p(X)$. We may assume that the points x_i are representable by shortest paths γ_i emanating from p.

By the definition of angle, $\tilde{\angle}\gamma_i(t)p\gamma_j(t) \to \angle(\gamma_i, \gamma_j) = |x_i x_j|$ as $t \to \infty$. Therefore one can find such a small $r > 0$ that $\tilde{\angle}\gamma_i(t)p\gamma_j(t) > \frac{1}{2}|x_i x_j| > \varepsilon/2$, and hence $|\gamma_i \gamma_j| > 2r\sin(\varepsilon/4) > r\varepsilon/4$, for all i, j $(i \neq j)$. In other words, we obtain an $(r\varepsilon/4)$-separated set in an r-ball. Applying Lemma 10.9.2 yields a desired bound for the number of points in the set. Since ε is arbitrary, it follows that $\Sigma_p(X)$ is compact. □

Recall that the tangent cone $K_p(X)$ over the space of directions $\Sigma_p(X)$ consists of the origin o and points represented by (ξ, r) where $x \in \Sigma_p(X)$ and $r > 0$. The distance from (ξ, r) to the origin equals r, and the distance from $a_1 = (\xi_1, r_1)$ and $a_2 = (\xi_2, r_2)$ is defined so that $\tilde{\angle}_0 a_1 o a_2 = \angle(\xi_1, \xi_2)$.

Theorem 10.9.3. *Let X be a finite-dimensional Alexandrov space and $p \in X$. Then the Gromov–Hausdorff tangent cone of X at p exists and is isometric to $K_p(X)$.*

Proof. Let B denote the unit ball centered at the origin o of the cone $K_p(X)$. We have to show that the rescaled balls $\frac{1}{r}B_r(p)$ converge to B in the Gromov–Hausdorff metric as $r \to 0$. Fix an $\varepsilon > 0$. By one of the criteria of Gromov–Hausdorff convergence (namely, by Proposition 7.4.11), it suffices to find finite ε-nets $\{x_i\}$ in B and $\{y_i\}$ in $\frac{1}{r}B_r(p)$, where r is small enough, such that $\big||x_ix_j| - |y_iy_j|\big| < \varepsilon$ for all i, j.

Pick a finite ε-net $\{x_i\}$ in B whose elements are representable by shortest paths. Let γ_i denote a shortest path representing x_i, parameterized with the constant speed $|ox_i|$. Assume that r is so small that the point $p_i\gamma_i(r)$ is defined for every i, and

$$\left|\frac{|p_ip_j|}{r} - |x_ix_j|\right| < \varepsilon$$

for all i, j. (This inequality for a small r is an immediate consequence of the definition of the tangent cone.) Now let $\{y_i\}$ be the points corresponding to $\{p_i\}$ in the rescaled ball $\frac{1}{r}B_r(X)$. The above inequality means that $\big||y_iy_j| - |x_ix_j|\big| < \varepsilon$ for all i, j.

It remains to show that $\{y_i\}$ is an ε-net in $\frac{1}{r}B_r(X)$, or, equivalently, in the unit ball of the rescaled space $X' = \frac{1}{r}X$. To do this, recall that the "logarithm" map $\log_p : X' \to K_p(X') = K_p(X)$ from Lemma 10.6.11 is noncontracting if X is of curvature ≥ 0. Since $\{x_i\}$ is an ε-net in the unit ball of $K_p(X)$, for any point y in the unit ball of X' there is an i such that $|\log_p(y)x_i| < \varepsilon$. Since $\log_p$ is noncontracting and we may assume that $x_i = \log_p(y_i)$, it follows that $|yy_i| \leq \varepsilon$. Therefore $\{y_i\}$ is an ε-net.

If the curvature bound k is nonzero, the same argument works with minor corrections. Since the rescaled space $X' = \frac{1}{r}X$ has curvature $\geq kr^2$ (which goes to zero as $r \to 0$), for a sufficiently small r the logarithm map satisfies $|\log_p(x)\log_p(y)| \geq \frac{1}{2}|xy|$ within the unit ball. Then it follows that $\{y_i\}$ is a (2ε)-net. □

Remark 10.9.4. Note that the above argument not only proves that $K_p(X)$ is the Gromov–Hausdorff limit of rescaled balls, but also gives explicit small-distortion maps from the balls to the cone. Namely the distortion of the "logarithm" map goes to zero with the radius of the ball. Reformulating this in terms of angles instead of distances, one gets the following useful fact: for any $\varepsilon > 0$ there is an $r > 0$ such that $|\widetilde{\angle}xpy - \angle xpy| < \varepsilon$ whenever $|px|$ and $|py|$ are less than r.

Corollary 10.9.5. *Let X be an n-dimensional Alexandrov space and $p \in X$. Then the tangent cone $K_p(X)$ is an n-dimensional Alexandrov space of nonnegative curvature.*

Proof. $K_p(X)$ is a length space and has nonnegative curvature as a Gromov–Hausdorff limit (by Proposition 10.7.1). More precisely, if X has curvature $\geq k$, then $\lambda^{-1}X$ has curvature $\geq \lambda^2 k$, and this curvature bound goes to 0 as $\lambda \to 0$. Then the Gromov–Hausdorff limit $K_p(X)$ has curvature $\geq -\varepsilon$ for any $\varepsilon > 0$, and hence it is nonnegatively curved.

It remains to prove that $\dim_H(K_p(X)) = \dim_H(X)$. The inequality $\dim_H(K_p(X)) \geq \dim_H(X)$ follows from the fact that there is a noncontracting map from X to $K_p(X)$, namely, the "logarithm" map (see Lemma 10.6.11). The inverse inequality follows from Theorem 10.9.3 and Corollary 10.8.25 about the dimension of a Gromov–Hausdorff limit. □

Corollary 10.9.6. *Let X be an n-dimensional Alexandrov space and $p \in X$. Then the space of directions $\Sigma_p(X)$ is an $(n-1)$-dimensional Alexandrov space of curvature ≥ 1, provided that $n \geq 2$. If $n = 1$, then $\Sigma_p(X)$ consists of one or two points.*

Proof. Since the cone $K_p(X)$ is nonnegatively curved, Theorem 10.2.3 implies that $\Sigma_p(X)$ is a length space of curvature ≥ 1 or a pair of points. (The additional requirement that no triangle in $\Sigma_p(X)$ has perimeter greater than 2π follows from Exercise 10.1.3.) Since $\dim_H K_p(X) = n$, one has $\dim_H \Sigma_p(X) = n - 1$. This seemingly obvious implication is left as an exercise to the reader. □

10.9.2. Regular points.

Definition 10.9.7. Let X be an n-dimensional Alexandrov space. A point $p \in X$ is said to be *regular* if the space of directions $\Sigma_p(X)$ is isometric to the standard sphere S^{n-1}.

Equivalently, p is regular if the tangent cone $K_p(X)$ is isometric to $\mathbb{R}^n$.

Example 10.9.8. If X is a convex polyhedral surface in $\mathbb{R}^3$, then all points but vertices are regular. This is because the intrinsic metric of a polyhedral surface is flat everywhere except at vertices.

Example 10.9.9. If X is the doubly-covered disc (that is, two copies of the disc glued along the boundary), then all points of X are regular.

The following lemma is an interesting global fact about Alexandrov spaces of curvature ≥ 1. It tells that one can recognize the standard sphere just by pointing out a finite collection of points with certain distances between them.

Lemma 10.9.10. *Let Y be an Alexandrov space of curvature ≥ 1 and of dimension no greater than $n-1$ $(n \geq 2)$. Suppose that there are n pairs of points $\{(x_i, y_i)\}_{i=1}^n$ in Y such that $|x_i y_i| = \pi$ for all i, and $|x_i x_j| = |x_i y_j| = |y_i y_j| = \frac{\pi}{2}$ if $i \neq j$. Then Y is isometric to S^{n-1}.*

We refer to a collection $\{(x_i, y_i)\}$ satisfying the assumptions of the lemma as an *orthogonal collection*. We mainly use this term in the case when Y is a space of directions. In this case, the lemma tells us that a point in an n-dimensional Alexandrov space is regular if and only if there is an orthogonal collection of n pairs of directions at this point.

Proof. We prove the lemma via an induction on n. Since $\operatorname{diam}(Y) = \pi$, Y is a suspension (spherical cone) over an Alexandrov space Z of curvature ≥ 1 (see Exercise 10.4.3). Moreover one can take the pair (x_n, y_n) as the poles of the suspension; then

$$Z = \{x \in Y : |xx_n| = |xy_n| = \frac{\pi}{2}\}.$$

In particular, Z contains the points x_i and y_i for all $i = 1, \dots, n-1$. Thus Z is isometric to S^{n-2} by the inductive assumption and therefore Y is isometric to S^{n-1}. □

Exercise 10.9.11. Let Y be an Alexandrov space of curvature ≥ 1 containing n pairs of points $\{(x_i, y_i)\}$ such that $|x_i y_i| = \pi$ for all i and the determinant of the $n \times n$ matrix $(\cos |x_i x_j|)$ is nonzero. Prove that Y contains a subset isometric to the standard sphere S^{n-1}, and that Y is isometric to S^{n-1} if $\dim_H(Y) = n - 1$.

Corollary 10.9.12. *A point p in an n-dimensional Alexandrov space is regular if and only if p is (n, ε)-strained for any $\varepsilon > 0$.*

Proof. Suppose that p is regular. Then there exist n pairs of directions $\{(\xi_i, \eta_i)\}_{i=1}^{\infty}$ in $\Sigma_p(X)$ with $\measuredangle(\xi_i, \eta_i) = \pi$ for all i and $\measuredangle(\xi_i, \xi_j) = \measuredangle(\xi_i, \eta_j) = \measuredangle(\eta_i, \eta_j) = \pi/2$ for $i \neq j$. Consider shortest paths whose directions are close to ξ_i and η_i and take points close to p on these shortest path. These points form an (n, ε)-strainer (see the third part of Proposition 10.8.12).

Now suppose that p is (n, ε)-strained for any $\varepsilon > 0$. Then for any $\varepsilon > 0$ the space of directions $\Sigma_p(X)$ contains n pairs of points (ξ_i, η_i) such that $|\xi \eta_i| > \pi - \varepsilon$, and other distances between these points are (11ε)-close to $\frac{\pi}{2}$ (see the second part of Proposition 10.8.12). Since $\Sigma_p(X)$ is compact, one can extract a converging sequence of such pairs corresponding to a sequence $\varepsilon_i \to 0$. The limit is a collection of points in Σ_p satisfying the conditions of Lemma 10.9.10; hence $\Sigma_p(X)$ is isometric to S^{n-1}. □

Corollary 10.9.13. *The set of regular points in an Alexandrov space is dense and moreover is an intersection of a countable collection of open dense sets.*

Proof. Let X be an n-dimensional Alexandrov space. For every positive integer i consider the set S_i of $(n, 1/i)$-strained points. By Corollary 10.8.24,

S_i is an open dense set. By Corollary 10.9.12, the set of regular points is the intersection of all S_i. Now Baire's Theorem 1.5.13 implies that this set is dense in X. □

Our next goal is to prove that a small neighborhood of a regular point is Lipschitz-close to a Euclidean region. We begin with a refinement of Lemma 10.9.10. It tells us that there is a certain stability in Lemma 10.9.10; namely, if a space of curvature ≥ 1 contains an "almost orthogonal" collection of directions, then it is close to the sphere in the Gromov–Hausdorff topology.

This proof illustrates the important technique of using Gromov's Compactness Theorem. This technique allows us to derive "stability" statements from "rigidity" statements almost automatically in many cases; see for example Exercise 10.9.15 below.

Lemma 10.9.14. *For every integer $n \geq 2$ and every $\varepsilon > 0$ there is $\delta > 0$ such that the following holds.*

Let Y be an Alexandrov space of curvature ≥ 1 and of dimension no greater than $n-1$. Suppose that Y contains n pairs of points $\{(x_i, y_i)\}_{i=1}^n$ such that $|x_i y_i| > \pi - \delta$ for all i, and the distances $|x_i x_j|$, $|x_i y_j|$, $|y_i y_j|$ $(i \neq j)$ are δ-close to $\pi/2$. Then $d_{GH}(Y, S^{n-1}) < \varepsilon$.

Proof. Suppose the contrary. Then there is a sequence of spaces $\{Y_m\}_{m=1}^\infty$ satisfying the assumptions of the lemma for $\delta = \delta_m$ where $\delta_m \to 0$ and such that $d_{GH}(Y_m, S^{n-1}) \geq \varepsilon$. Due to the Compactness Theorem 10.7.2 we may assume that $\{Y_m\}$ converges in the Gromov–Hausdorff topology to an Alexandrov space Y of curvature ≥ 1 and of dimension no greater than $n-1$. We are going to show that Y is isometric to S^{n-1}, contrary to our assumption.

Recall (Theorem 7.3.25) that there is a correspondence between Y_m and Y with distortion no greater than $2d_{GH}(Y_m, Y)$. Thus an "almost orthogonal" collection of points in Y_m corresponds to an "almost orthogonal" collection in Y (with δ_m replaced by $\delta_m + d_{GH}(Y_m, Y)$ which goes to zero anyway). Extracting a converging subsequence of these collections in Y, one obtains an orthogonal collection. By Lemma 10.9.10, it follows that $Y \simeq S^{n-1}$. □

Here are a few exercises on similar use of the Compactness Theorem.

Exercise 10.9.15. Prove that for any $n \in N$ and $\varepsilon > 0$ there is a $\delta > 0$ such that for any n-dimensional Alexandrov space X of curvature ≥ 1 the following holds.

1. If $\operatorname{diam}(X) > \pi - \delta$, then X is ε-close in the Gromov–Hausdorff metric to a spherical cone over some Alexandrov space of curvature ≥ 1 and dimension at most $n-1$. (Compare with Exercise 10.4.3.)

2. If the radius of X (see Exercise 10.4.5) is greater than $\pi - \delta$, then X is ε-close to S^n, i.e., $d_{GH}(X, S^n) < \varepsilon$.

3. If $\mu_n(X) > \mu_n(S^n) - \delta$, then X is ε-close to S^n.

Hint: Use the Bishop inequality to reduce this statement to the previous one.

Theorem 10.9.16. *For any integer $n \geq 1$ and every $\varepsilon > 0$ there is a $\delta > 0$ such that any (n, δ)-strained point in any n-dimensional Alexandrov space has a neighborhood which is ε-close with respect to the Lipschitz distance (cf. Section 7.2) to an open region in $\mathbb{R}^n$.*

We give two slightly different proofs demonstrating different techniques.

Proof. We will show that, in a sufficiently small neighborhood U, the map $f : U \to \mathbb{R}^n$ of distance coordinates associated with an (n, δ)-strainer $\{(a_i, b_i)\}$, and the inverse map f^{-1} have Lipschitz constants not greater than $1 + \varepsilon$.

We may assume that $\{(a_i, b_i)\}$ is an (n, δ)-strainer for any point of U. We want to show that for any two points $x, y \in U$ the ratio $|f(x)f(y)|/|xy|$ is close to 1. Considering a comparison triangle for $\triangle a_i xy$, one sees that

$$f_i(y) - f_i(x) = |a_i y| - |a_i x| = |xy| \cdot \cos \tilde{\angle} a_i xy + o(|xy|).$$

Then

$$|f(y) - f(x)|^2 = \sum |f_i(y) - f_i(x)|^2 = |xy|^2 \cdot \sum \cos^2 \tilde{\angle} a_i xy + o(|xy|^2).$$

To prove that the ratio $|f(x) - f(y)|/|xy|$ is close to 1, it suffices to show that $\sum \cos^2 \tilde{\angle} a_i xy$ is close to 1. Fix shortest paths connecting x to y, a_i and b_i, and denote by ξ, ξ_i and η_i the respective elements of the space of directions $\Sigma_x(X)$. By the last statement of Lemma 10.8.13, the angles $|\xi\xi_i| = \angle a_i xy$ between these shortest paths are almost equal to the comparison angles $\tilde{\angle} a_i xy$. Therefore it is sufficient to prove that $\sum \cos^2 |\xi\xi_i|$ is close to 1.

Note that this sum equals 1 if the space of directions $\Sigma_x(X)$ is the standard sphere S^{n-1} and $\{(\xi_i, \eta_i)\}$ is an orthogonal collection (in the sense of Lemma 10.9.10). What we know is that $\{(\xi_i, \eta_i)\}$ is "almost orthogonal" (up to a term of order δ), and hence by Lemma 10.9.14 the space of directions $\Sigma_x(X)$ is close to S^{n-1} in the Gromov–Hausdorff metric. Then by Theorem 7.3.25 there is a correspondence with a small distortion between $\Sigma_x(X)$ and S^{n-1}. Let $\bar{\xi}$, $\bar{\xi}_i$ and $\bar{\eta}_i$ be points in the sphere corresponding to ξ, ξ_i and η_i. Then

$$\sum \cos^2 |\xi\xi_i| \approx \sum \cos^2 |\bar{\xi}\bar{\xi}_i| \approx 1$$

(the second approximate equality follows from the fact that in the sphere there is a real orthogonal collection near $\{(\bar{\xi}_i, \bar{\eta}_i)\}$. □

Another proof. One can prove the theorem using regular points instead of Lemma 10.9.14. Namely, the above argument shows that the approximate equality $\sum \cos^2 |\xi\xi_i| \approx 1$ holds if $\Sigma_x(X)$ is isometric to S^{n-1}, i.e., x is a regular point. Then the inequalities $(1+\varepsilon)^{-1}|xy| \le |f(x)f(y)| \le (1+\varepsilon)|xy|$ follow for a regular point x. Since regular points are dense, these inequalities follow for all x, y. □

Note that the second proof makes it possible to write an explicit bound for δ in terms of ε, while the first one only proves that a δ exists (see how Lemma 10.9.14 was proved).

Sometimes the following reformulation of Theorem 10.9.16 is useful.

Corollary 10.9.17. *For any integer $n \ge 1$ and every $\delta > 0$ there is a $\varepsilon > 0$ such that the following holds. If X is an n-dimensional Alexandrov space and $p \in X$ is a point such that $d_{GH}(\Sigma_p(X), S^{n-1}) < \varepsilon$, then p has a neighborhood which is δ-close with respect to the Lipschitz distance to an open region in $\mathbb{R}^n$.*

Remark 10.9.18. Points $p \in X$ with $d_{GH}(\Sigma_p(X), S^{n-1}) < \varepsilon$ are referred to as *ε-regular points*. The set X_ε of ε-regular points features some interesting properties. Obviously every point of X_ε is $(n, 2\varepsilon)$-strained and hence has a Euclidean neighborhood. For the same reason, the interior of X_ε is dense in X (see Corollary 10.8.23).

Furthermore, though it sounds a little surprising, the set X_ε is convex in the sense that a shortest path connecting two points of X_ε does not leave X_ε. This follows from the fact that the tangent cone is constant along any shortest path except its endpoints; in other words, if x and y belong to the interior of some shortest path, then $K_x(X)$ is isometric to $K_y(X)$ [**Pet**].

Corollary 10.9.19. *Let X be an n-dimensional Alexandrov space and $p \in X$ be a regular point. Then p has neighborhoods arbitrarily close to regions in $\mathbb{R}^n$ in the sense of Lipschitz distance.*

Exercise 10.9.20. Let $n \in N$, $k \in \mathbb{R}$ and $r > 0$ be fixed. Prove that there is a $\delta > 0$ such that the following holds. If X is an n-dimensional Alexandrov space of curvature $\ge k$, $p \in X$ and $\mu_n(B_r(p)) > V^n_{k,r} - \delta$, where $V^n_{k,r}$ is the volume of the r-ball in the n-dimensional space form of curvature k, then p has a neighborhood homeomorphic to $\mathbb{R}^n$.

In particular, if the volume of any r-ball in X is no less than $V^n_{k,r} - \delta$, then X is a manifold.

Hint: Using a modified argument from the proof of the Bishop inequality, show that the $(n-1)$-dimensional measure of $\Sigma_p(X)$ is close to that of S^{n-1}. Then apply the last part of Exercise 10.9.15.

10.10. Further Information

10.10.1. Stratification. One says that a collection $\{X_i\}_{i=1}^N$ of subsets of a topological space X forms a (finite) *stratification* of X into topological manifolds if

(1) The sets X_i are disjoint, and $\bigcup_{i=1}^N X_i = X$.
(2) Every set X_i is a topological manifold (without boundary).
(3) $\dim X_1 > \dim X_2 > \cdots > \dim X_N$.
(4) For every $k = 1, \ldots, N$, the set $X_k^+ = \bigcup_{i=k}^N X_i$ is closed in X. Equivalently, the closure of X_k is contained in X_k^+ for all k.

The sets X_i are called *strata*. Each stratum is not necessarily connected.

Theorem 10.10.1. *Every finite-dimensional Alexandrov space X admits a stratification into topological manifolds.*

Moreover, there is a canonical stratification such that every component C of every stratum is topologically homogeneous in the following sense: every two points $p, q \in C$ have neighborhoods in X which are pointed homeomorphic to each other (pointed means that the homeomorphism maps p to q).

This stratification can be described as follows: the stratum X_1 is the set of points possessing a Euclidean neighborhood in X; then X_2 is the set of points in $X \setminus X_1$ possessing a Euclidean neighborhood of maximal possible dimension, and so on. The matter is that this procedure exhausts the whole space. For example, a disc is stratified into its interior and its boundary; the cone over $\mathbf{RP}^2$ is stratified into the complement of the origin (which is a three-dimensional flat space) and the origin itself.

Corollary 10.9.6 allows us to give an inductive definition of the *boundary* of an Alexandrov space. First, the boundary of a one-dimensional space is its topological boundary (recall that all one-dimensional Alexandrov spaces are segments, rays, lines and circles). Then, assuming that the notion of boundary is already defined in dimensions $1, \ldots, n-1$, define the boundary of an n-dimensional Alexandrov space X as the set of points $p \in X$ such that the space of directions $\Sigma_p(X)$ has nonempty boundary.

Using the above stratification theorem, it is possible to prove that the boundary of an Alexandrov space possesses various nice properties; for

example, it is always a closed subset of codimension one. In fact, the boundary coincides with the closure of the stratum of codimension one.

It can be proved that there is no stratum of codimension two. So if an Alexandrov space has no boundary the set of topologically singular points is very thin: it has codimension at least three.

The following theorem is a local counterpart of the Theorem 10.10.1.

Theorem 10.10.2. *Every point p of a finite-dimensional Alexandrov space has a neighborhood pointed homeomorphic to the tangent cone at p.*

Though Theorem 10.10.2 looks much simpler than Theorem 10.10.1, there is no known "simple" proof for it. Both theorems can be proved simultaneously by induction in dimension, and the proof of the second theorem in dimension n involves the first theorem in dimensions less than n.

The proof of both theorems (and many other strong results as well) is contained in the manuscript "A.D.Alexandrov's spaces with curvatures bounded from below, II" by G. Perelman. Unfortunately this paper is not published.

Actually Theorem 10.10.1 can be formulated in a stronger form. Namely, there is a canonical stratification into so-called *primitive extremal* subsets. This is proved in [**Per1**].

We do not define here what is an extremal set. Let us only mention that the boundary of an Alexandrov space is always an extremal set. Another example of an extremal set is a point whose space of directions has diameter less than π.

The proofs of the above mentioned theorems are complicated and are based on study of analytical properties of distance functions and some nonelementary facts from topology. These topics are beyond the scope of this book.

In dimension 2, the results about topological structure of Alexandrov spaces take very simple form:

Corollary 10.10.3. *Every two-dimensional Alexandrov space is a topological manifold, possibly with boundary.*

Exercise 10.10.4. 1. Prove Corollary 10.10.3 directly (without the stratification theorem).

2. Let X be a three-dimensional Alexandrov space having no boundary. Prove that topologically singular points of X are isolated and, for every such point p, the space of directions $\Sigma_p(X)$ is homeomorphic to $\mathbf{RP}^2$.

The reader can observe that a classification of topological singularities in dimension n is roughly equivalent to a classification of topological types

of Alexandrov spaces of curvature ≥ 1 in dimension $n-1$. This appears easy for $n \leq 3$ (see the exercise above) but in general to obtain such a classification seems hopeless.

Generalizing Theorem 10.10.2 above, a very strong *stability theorem* is proved in the same manuscript by G. Perelman.

Theorem 10.10.5 (Stability theorem). *For every $k \in \mathbb{R}$ and for every compact Alexandrov space X^n of curvature $\geq k$, there exists an $\varepsilon > 0$ such that every compact Alexandrov space Y^n of curvature $\geq k$ and such that*

$$d_{GH}(X^n, Y^n) < \varepsilon$$

is homeomorphic to X^n (here d_{GH} is the Gromov–Hausdorff distance and upper indices denote the dimension of spaces).

In particular, the theorem implies that an Alexandrov space having topological singularities cannot be approximated by Riemannian manifolds of the same dimension and with the same curvature bound.

10.10.2. Convexity arguments. Let X be a locally compact length space. Recall that a function $f : X \to \mathbb{R}$ is said to be convex if its restriction to every geodesic is a convex function in one variable.

Theorem 10.10.6. *Let X be a finite-dimensional Alexandrov space of nonnegative curvature with nonempty boundary. Then the distance function $d(\cdot, \partial X)$ is convex. Moreover it is strictly convex if X has curvature $\geq k > 0$.*

Clearly a convex subset of an Alexandrov space is an Alexandrov space with the same curvature bound. In particular, such subsets have well-defined boundary. Jointly with the theorem above, this allows one to extend the proof of the famous *Soul Theorem* by J. Cheeger and D. Gromoll [**CG1**] to Alexandrov spaces.

Theorem 10.10.7. *Let X^n be a noncompact finite-dimensional Alexandrov space of nonnegative curvature. Then there is a convex compact subset $S \subset X^n$ without boundary and such that S is a deformation retract of X^n.*

Note that if X is a Riemannian manifold, a convex closed subset without boundary is a totally geodesic submanifold.

Remark 10.10.8. In the case of Riemannian manifolds, the Soul Theorem can be formulated in a stronger form: X is homeomorphic (in fact, diffeomorphic) to the total space of the normal bundle of S in X. For Alexandrov spaces a similar strengthening is false as the following example shows:

Example 10.10.9. (G. Perelman) The quotient of the complex space $\mathbb{C}^k$ under the equivalence $x \sim \lambda x$, $\lambda \in \mathbb{C}$, $|\lambda| = 1$, is the cone over the complex projective space $\mathbb{CP}^k$; denote this quotient by $K_0(\mathbb{CP}^k)$.

The natural projection $(z_1, \ldots, z_3) \to (z_1, z_2)$ of $\mathbb{C}^3$ onto $\mathbb{C}^2$ induces the projection $\pi : K_0(\mathbb{CP}^2) \to K_0(\mathbb{CP}^1)$. Let us equip $\mathbb{CP}^k$ with the canonical metric having sectional curvatures between 1 and 4, and equip the cones with corresponding cone metrics. Let $\overline{B}_0(1)$ be the unit ball in $\mathbb{CP}^1$ centered at the origin. Set $X^5 = \pi^{-1}(\overline{B}_0(1))$. X^5 is a convex subset of $\mathbb{CP}^2$, and therefore it is a complete noncompact Alexandrov space of nonnegative curvature and with nonempty boundary. It is possible to prove that the doubling $\overline{X}^5$ of X^5 is a space of the same type but without boundary. Since 0 is a topologically singular point of X^5, the space $\overline{X}^5$ has two singular points. One can show that the doubling S of $\overline{B}_0(1)$ is a soul of $\overline{X}^5$. Since S is homeomorphic to a sphere S^3, a fiber bundle over S is a manifold and hence cannot be homeomorphic to $\overline{X}^5$.

10.10.3. Hausdorff measure. Recall that the m-dimensional Hausdorff measure $\mu_m(X)$ of m-dimensional compact Alexandrov space X is always finite and positive. And as we already mentioned after Theorem 10.8.5, the set of nonregular points has zero measure; in particular, the set of topological singularities has zero measure. Moreover, for every sufficiently small $\varepsilon > 0$ (say, $\varepsilon < \frac{1}{100m}$) one has $\mu_m(X \setminus X_\varepsilon) = 0$, where X_ε is the set of ε-regular points.

Let $\{X_i\}$ be a sequence of compact metric spaces, $X_i \xrightarrow[GH]{} X$ as $i \to \infty$. Then one can choose a sequence of numbers ν_i such that $\nu_i \to 0$ as $i \to \infty$ and X_i is a ν_i-approximation for X. By Corollary 7.3.28 there exist maps $f_i : X \to X_i$ such that $\operatorname{dis} f_i < 2\nu_i$ and $f_i(X)$ is $2\nu_i$-net in X_i. One says that measures ω_i defined on X_i *weakly converge* to a measure ω on X if for every closed set $E \in X$ such that $\omega(\partial E) = 0$ one has $\omega_i(U_{2\nu_i}(f_i^{-1}(E))) \to \omega(E)$ as $i \to \infty$.

Theorem 10.10.10. *Let m-dimensional compact Alexandrov spaces X and X_i, $i = 1, 2, \ldots$, have the same curvature bound and $X_i \xrightarrow[GH]{} X$ as $i \to \infty$. Then m-dimensional Hausdorff measures of X_i converge weakly to that of X (for any choice of f_i and ν_i in the definition of weak convergence).*

This theorem has an obvious

Corollary 10.10.11. *Let X_i be an m-dimensional compact Alexandrov space with the same curvature bound and converging to a compact space X. Then $\mu_m(X_i) \to 0$ if and only if $\dim_H X < m$.*

10.10.4. Collapse. The situation described in the previous corollary is called a *collapse*. More precisely, one says that a sequence $\{X_i\}$ of m-dimensional Alexandrov spaces with the same curvature bound *collapses* to a compact space Y if $X_i \xrightarrow[GH]{} Y$ and $\mu_m X_i \to 0$ as $i \to \infty$ (or, equivalently, $\dim Y < m$).

It is clear that in this case Y is an Alexandrov space with the same curvature bound. Let us restrict ourselves to a partial case when the spaces X_i are manifolds homeomorphic to one another or, equivalently, that we have a sequence of metrics on a fixed manifold.

Then, if the limit space Y is a Riemannian manifold or at least X has only weak singularities, then X_i (for all sufficiently large i) admits a structure of locally trivial fibration over Y (see [**Yam**]).

If a collapse happens for a sequence of manifolds having in addition a common *upper* curvature bound, then the fibers of this fibration have a very special topological type; namely, they are so-called nil-manifolds (see [**CFG**]).

Another interesting question concerning collapse is whether a given Alexandrov space can be approximated by Riemannian manifolds (with some fixed lower curvature bound). Recall that due to Stability Theorem 10.10.5, an Alexandrov space X having a topological singularity cannot be approximated by Riemannian manifolds of the same dimension. This does not exclude approximation by higher dimensional manifolds, i.e., with collapse. There is an example of an Alexandrov space of nonnegative curvature which cannot be approximated (even with collapse) by nonnegatively curved Riemannian manifolds; this however does not exclude approximation by manifolds having a bit smaller curvature bounds. No affirmative results of this sort are known, and the problem looks very hard.

10.10.5. Quasigeodesics. Geodesic lines in a smooth surface or a Riemannian manifold can be characterized in two ways: they are locally shortest paths, and they are straightest paths. The latter means that geodesics are curves of zero geodesic curvature. Geodesics in a length space are defined as locally shortest paths. Unfortunately geodesics in Alexandrov spaces lack some useful properties of Riemannian geodesics. First, there may be points such that no geodesic passes through them (for instance, consider the apex of a sharp 2-cone). Even at a regular point, there may be directions not representable by shortest paths (emanating from the point). And not every geodesic can be extended beyond its endpoint even if this point is regular. These circumstances stimulate consideration of a wider class of curves, so-called *quasigeodesics*. Roughly speaking, quasigeodesics are locally straightest paths in Alexandrov spaces.

To give a formal definition, recall the study of distance functions in Subsection 4.1.2. Given a point p and a geodesic γ in a length space X, we considered the distance function $g(t) = d(\gamma(t), p)$ and compared it with the similar distance function g_k (with the same values at endpoints) defined by a point and a geodesic in the k-plane. The definition of a space of curvature

$\geq k$ says that for every sufficiently short segment of γ and every point p sufficiently close to it, the function g must be no less concave than g_k.

The same comparison property may hold for curves γ that are not geodesics. Curves possessing this property are called k-quasigeodesics. A *quasigeodesic* is a k-quasigeodesic in an Alexandrov space of curvature $\geq k$.

Quasigeodesics have many nice properties; in particular all the above mentioned missing properties of geodesics do hold for quasigeodesics.[3]

In addition, the class of quasigeodesics is closed under noncollapsing Gromov–Hausdorff convergence.

Let us mention one more property. If X is a space of curvature ≥ 1, two points $a, b \in X$ are called antipodes if $|ax| + |xb| \leq \pi$ for all $x \in X$. If two quasigeodesics have a common endpoint p and their directions at p are antipodes, then the concatenation of these paths (taken with the opposite orientation) is a quasigeodesic too. This allows one to build quasigeodesics "from pieces".

A geodesic in an extremal set (in particular, a geodesic in the boundary) is a quasigeodesic in the ambient Alexandrov space [**PP**].

There are several important applications of quasigeodesics. One of them is the proof of the following gluing theorem (see [**Pet1**] and the preprint mentioned above).

Theorem 10.10.12. *Let X and Y be Alexandrov spaces of the same dimensions and of curvature $\geq k$. Suppose that their boundaries are isometric. Then the space Z obtained by gluing X and Y along an isometry of their boundaries is a space of curvature $\geq k$.*

10.10.6. Analysis on Alexandrov spaces. Recall that almost every point of an Alexandrov space X has a bi-Lipschitz Euclidean neighborhood. So one can consider "atlases" of bi-Lipschitz charts covering X up to a set of zero measure. In such a chart, a Riemannian metric tensor can be defined at almost every point (using the fact that a Lipschitz function is differentiable almost everywhere, and that the tangent cone of X is Euclidean at almost every point). The coefficients of such a metric tensor are *a priori* only measurable functions, and their relation to metric properties of X are not clear. One may ask if it is possible to choose local coordinates so that (some of) the machinery of differential geometry works. For example, one may want to define Christoffel symbols (as functions, or measures, or whatever), use Riemannian formulas for first and second variation, and so on. To make this analysis work, one has to find coordinates in which the metric coefficients

[3]See the preprint by G. Perelman and A. Petrunin, *Quasigeodesics and gradient curves in Alexandrov spaces.*

are as regular as possible (at least, differentiable almost everywhere), and/or utilize some generalized interpretation of differential formulas.

There are some affirmative results in this direction. Let us mention [**OS**] and [**Ot**] among already published papers; several more papers are to appear soon. (This field is far from being complete and is changing rapidly.)

10.10.7. Curvature bounded from both sides. One may wonder what can be said about a space of curvature bounded both above and below. The answer is simple though it required a hard work to prove it: loosely speaking, such a space is a Riemannian manifold whose metric tensor has slightly lower smoothness than usually considered in Riemannian geometry. More precisely, the following theorem holds.

Theorem 10.10.13 (I. Nikolaev, see [**BN**]). *Let (X, d) be a length space with both side upper and lower curvature bounds (of curvature $\geq k$ and $\leq K$). If (X, d) has no boundary, then X is a manifold and possesses a C^3-smooth atlas such that in its charts the metric d can be defined by a Riemannian metric tensor whose coefficients g_{ij} are in the class W_p^2 for every $p > 1$.*

If X has a boundary, then the interior of X is an almost Riemannian manifold in the same sense as above; however the boundary may be non-smooth.

If $k = K$, then (X, d) is usual space of constant curvature.

Recall that W_p^2 means the space of functions whose second Sobolev derivatives belong to L_p. (In particular $g_{ij} \in C^{1,\alpha}$ for all $\alpha < 1$.)

The theorem allows us to define formal Christoffel symbols and curvature tensor via derivatives of g_{ij}. These formal values are defined almost everywhere; nevertheless they have a reasonable geometric meaning. For instance, parallel transport defined geometrically coincides (for almost all paths) with one given by Christoffel symbols. And formal sectional curvatures have geometric meaning as well.

Also I. Nikolaev proved that such a space can be approximated (in the sense of uniform convergence) by Riemannian manifolds of the same dimension and with curvature bounds converging to that of (X, d).

Bibliography

[AB] S. Alexander, R. Bishop, *The Hadamard-Cartan theorem in locally convex metric spaces*, Enseign. Math. (2) **36** (1990), no. 3-4, 306-320

[AB1] S. Alexander, R. Bishop, *Warped products of Hadamard spaces*, Manuscripta Math., **96** (1998), 487-505

[Al] A. D. Alexandrov, *Intrinsic Geometry of Convex Surfaces* (in Russian). German translation: *Die innere Geometrie der konvexen Flächen*, Akad. Verl., Berlin, 1955.

[Bab] I.K.Babenko, *Forte souplesse intersystolique de variétés fermées et de polyèdres*, Annales de l'Ecole Norm. Super. (to appear).

[Bal] W. Ballmann, *Lectures on spaces of nonpositive curvature*, DMV Seminar **25**, Birkhäuser (1995), viii+112.

[Ban] V. Bangert *Minimal geodesics*, Erg. Theory and Dyn. Syst. **10** (1990), 263–286.

[BC] R. Bishop and R. Crittenden, *Geometry of Manifolds*, Academic Press, New York – London, 1964.

[BGS] W. Ballmann, M. Gromov, and V. Schroeder, *Manifolds of nonpositive curvature*, Progress in Mathematics **61**, Birkhäuser, Boston (1985), vi+263.

[BH] M. R. Bridson and A. Haffliger *Metric spaces of non-positive curvature*, in Ser. A Series of Comprehensive Stadies in Mathematics, vol. 319, Springer-Verlag, Berlin, 1999.

[BN] V. Berestovskii and I. Nikolaev, *Multidimensional generalized Riemannian spaces*, in Geometry IV. Non-regular Riemannian geometry. Encyclopaedia of Mathematical Sciences, Springer-Verlag, Berlin, 1993, 165–244.

[BM] M. Bonk, *Quasi-geodesic segments and Gromov hyperbolic spaces*, Geom. Dedicata **62** (1996), no. 3, 281–298.

[BP] R. Benedetti and C. Petronio, *Lectures on Hyperbolic Geometry*, Springer, 1992

[BR] A. Belläche and J.-J.Risler (editors), *Sub-Riemannian Geometry*, Progress in Mathematics, vol. 144, Birkhäuser, 1996.

[Bur] D. Burago, *Periodic metrics*, Advances in Soviet Math. **9** (1992), 205-210.

[BI] D. Burago and S. Ivanov, *On asymptotic volume of tori*, Geom. Func. Anal. **5** (1995), 800–808

[BB] Yu. Burago, S. Buyalo, *Metrics of upper bounded curvature on 2-polyhedra. II*, St. Petersburg Math. J. **10** (1999), 619–650.

[BGP] Yu. Burago, M. Gromov, G. Perelman, *A. D. Alexandrov spaces with curvatures bounded below*, Russian Math. Surveys **47** (1992), 1–58.

[BZ] Yu. Burago and V. Zalgaller, *Geometric Inequalities*, A Series of Comprehensive Studies in Mathematics **285**, Springer-Verlag, Berlin, 1988.

[Car] M. P. do Carmo, *Riemannian Geometry*, Birkhäuser, Boston, 1992.

[CE] J. Cheeger and D. Ebin, *Comparison Theorems in Riemannian Geometry*, North Holland, New York, 1975.

[CFG] J. Cheeger, K. Fukaya, and M. Gromov *Nilpotent structures and invariant metrics on collapsed manifolds*, J. Amer. Math. Soc. **5** , no. 2,(1992) 327–372.

[CG1] J. Cheeger and D. Gromoll, *On the structure of complete manifolds of non-negative curvature*, Ann. of Math. **96** (1972), 413-443.

[CG] J. Cheeger and D. Gromoll, *The splitting theorem for manifolds of nonnegative Ricci curvature*, J.Differential Geometry **6** (1971-72), 119-128.

[Cha] I. Chavel, *Riemannian Geometry — a modern introduction*, Cambridge Univ. Press, Cambridge, 1993.

[Ch] , C. H. Chen *Warped products of metric spaces of curvature bounded from above*, Trans. Amer. Math. Soc. **351** (1999),4727-4740.

[E] P. B. Eberlein, *Geometry of Nonpositively Curved Manifolds*, The University of Chicago Press, Chicago and London, 1996.

[Fe] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, 1969.

[GHL] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian geometry* , Springer-Verleg, Berlin, 1990, xiv+284.

[Gro1] M. Gromov, *Filling Riemannian manifolds*, J. Differential Geometry **18** (1983), 1–147.

[Gro2] M. Gromov, *Systols and intersystolic Inequalities*, Actes de la table rande de géométrie differentielle en l'honeur de Marcel Berger. Collection SMF **1** (1996), 291–362.

[Gro3] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces*, Progress in Mathematics **152**, Birkhäuser, Boston, 1999.

[Grove] K. Grove, *Metric differential geometry*, Lecture Notes in Math. **1263** (1987), 171–227.

[GP] K. Grove and P. Petersen (Editors), *Comparison geometry*, Papers from the Special Year in Differential Geometry held in Berkeley, CA. 1993–94. Research Institute Publications, **30** (1997), Cambridge University Press, Cambridge, x+262 pp.

[HM] S. Hartman and J. Mikusiński,, *The theory of Lebesgue measure and integration*, International series of monographs on Pure and Applied Mathematics, vol. 15, 1961.

[HS] J. Hass and P. Scott, *Bounded 3-manifolds admits negatively curved metric with concave boundary*, J. Differential Geom. **40** (1994), no. 3, 449-459

[J] J. Jost, *Nonpositive curvature: Geometric and analytic Aspects*, Lectures in Mathematics ETH Birkhäuser, 1997.

[Ka] H. Karcher, *Riemannian comparison constructions*, in Global Differential Geometry (S. S. Chern ed.), MAA Studies in Mathematics **27** (1989), 170–222.

[Kl] B. Kleiner, *The local structure of length spaces with curvature bounded above*, Math. Z. **231** (1999), no. 3, 409–456.

[KS] M. Katz and A. Suciu, *Systolic freedom of loopspace*, (2000), 1-13, preprint.

[Mas] W. S. Massey, *Algebraic topology: an introduction*, Graduate Texts in Mathematics, vil. 56. Springer, New York, 1987

[Mi] J. Milnor, *Topology from the differentiable viewpoint*, Univ. Press of Virginia, 1965.

[Mo] F. Morgan, *Geometric measure theory: a beginner's guide*, Academic Press, 1988.

[Mun] J. R. Munkres, *Topology: a first course*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975. xvi+413

[Mun1] J. R. Munkres, *Analysis on manifolds*, Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA, 1991, xiv+366 pp.

[N] J. Nash, *The imbedding problem for Riemannian manifolds*, Ann. Math. **63** (1956), 20–63.

[OS] Y. Otsu and T. Shioya, *The Riemannian structure of Alexandrov spaces*, J. Differential Geom. **39** (1994), no. 3, 629–658.

[Ot] Y. Otsu, *Differential geometric aspects of Alexandrov spaces*, in Comparison Geometry, Math. Sci. Res. Inst. Publ **30** Cambridge Univ. Press, Cambridge (1997), 135–148.

[Per1] G. Perelman, *The Beginning of the Morse Theory on Alexandrov spaces*, St. Petersburg Math. Journ. **5** (1994), no. 1, 205-214.

[Pet] A. Petrunin, *Parallel transportation for Alexandrov spaces with curvature bounded below*, Geom. Funct. Analysis **8** (1998), no. 1, 123-148.

[Pet1] A. Petrunin, *Applications of quasigeodesics and gradiant curves*, in Comparison Geometry, Math. Sci. Res. Inst. Publ **30**, Cambridge Univ. Press, Cambridge (1997), 203-219

[Pl] C. Plaut, *Metric spaces of curvature* $\geq k$, in "A handbook of Geometric Topology" (to appear)

[PP] G. Perelman and A. Petrunin, *Extremal subset in Alexandrov spaces and a generalised Liberman theorem*, St. Petersburg Math. Journ. **5** (1994), no. 1, 215-227.

[PP1] G. Perelman and A. Petrunin, *Quasigeodesics and gradient curves in Alexandrov spaces*, Preprint (1994), 1-23.

[Resh] Yu. G. Reshetnyak, *Nonexpanding maps to spaces of curvature not greater than* K, Siberian. Matem. J. **9**, no. 4 (1968), 918-927 (in Russian).

[Sp] M. Spivak, *Calculus on manifolds. A modern approach to classical theorems of advanced calculus*, W. A. Benjamin, Inc., New York-Amsterdam 1965, xii+144 pp.

[St] L.Stoyanov, *An estimate from above of the number of periodic orbites for semi-dispersed billiards*, Comm. Math. Phys. **24** (1989), no. 2, 217-227.

[Yam] T. Yamaguchi, *A convergence theorem in the geometry of Alexandrov spaces*, Actes de la Table Ronde de Geometrie Differentielle (Luminy, 1992), 601–642, Semin. Congr., 1, Soc. Math. France, Paris, 1996.

Index

图字：01-2016-2527 号

度量几何学教程
Duliang Jihexue Jiaocheng

图书在版编目(CIP)数据

度量几何学教程 = A Course in Metric Geometry：英文 / (俄罗斯) 德米特里・布拉戈 (Dmitri Burago), (俄罗斯) 尤里・布拉戈 (Yuri Burago), (俄罗斯) 谢尔盖・伊万诺夫 (Sergei Ivanov) 著．—影印本．—北京：高等教育出版社，2017.1
ISBN 978-7-04-046908-0
Ⅰ. ①度… Ⅱ. ①德… ②尤… ③谢… Ⅲ. ①几何学—教材—英文 Ⅳ. ①O18
中国版本图书馆 CIP 数据核字 (2016) 第 280453 号

策划编辑　赵天夫　　责任编辑　赵天夫
封面设计　张申申　　责任印制　韩　刚

出版发行　高等教育出版社
社址　北京市西城区德外大街 4 号
邮政编码　100120
购书热线　010-58581118
咨询电话　400-810-0598
网址　http://www.hep.edu.cn
http://www.hep.com.cn
网上订购　http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
印刷　涿州市星河印刷有限公司

开本　787mm × 1092mm　1/16
印张　27.25
字数　610 千字
版次　2017 年 1 月第 1 版
印次　2017 年 1 月第 1 次印刷
定价　169.00 元

[物 料 号　46908-00]

美国数学会经典影印系列

1. **Lars V. Ahlfors**, Lectures on Quasiconformal Mappings, Second Edition, 978-7-04-047010-9
2. **Dmitri Burago, Yuri Burago, Sergei Ivanov**, A Course in Metric Geometry, 978-7-04-046908-0
3. **Tobias Holck Colding, William P. Minicozzi II**, A Course in Minimal Surfaces, 978-7-04-046911-0
4. **Javier Duoandikoetxea**, Fourier Analysis, 978-7-04-046901-1
5. **John P. D'Angelo**, An Introduction to Complex Analysis and Geometry, 978-7-04-046998-1
6. **Y. Eliashberg, N. Mishachev**, Introduction to the *h*-Principle, 978-7-04-046902-8
7. **Lawrence C. Evans**, Partial Differential Equations, Second Edition, 978-7-04-046935-6
8. **Robert E. Greene, Steven G. Krantz**,
   Function Theory of One Complex Variable, Third Edition, 978-7-04-046907-3
9. **Thomas A. Ivey, J. M. Landsberg**,
   Cartan for Beginners: Differential Geometry via Moving Frames and
   Exterior Differential Systems, 978-7-04-046917-2
10. **Jens Carsten Jantzen**, Representations of Algebraic Groups, Second Edition, 978-7-04-047008-6
11. **A. A. Kirillov**, Lectures on the Orbit Method, 978-7-04-046910-3
12. **Jean-Marie De Koninck, Armel Mercier**, 1001 Problems in Classical Number Theory, 978-7-04-046999-8
13. **Peter D. Lax, Lawrence Zalcman**, Complex Proofs of Real Theorems, 978-7-04-047000-0
14. **David A. Levin, Yuval Peres, Elizabeth L. Wilmer**, Markov Chains and Mixing Times, 978-7-04-046994-3
15. **Dusa McDuff, Dietmar Salamon**, *J*-holomorphic Curves and Symplectic Topology, 978-7-04-046993-6
16. **John von Neumann**, Invariant Measures, 978-7-04-046997-4
17. **R. Clark Robinson**,
   An Introduction to Dynamical Systems: Continuous and Discrete, Second Edition, 978-7-04-047009-3
18. **Terence Tao**,
   An Epsilon of Room, I: Real Analysis: pages from year three of a mathematical blog, 978-7-04-046900-4
19. **Terence Tao**, An Epsilon of Room, II: pages from year three of a mathematical blog, 978-7-04-046899-1
20. **Terence Tao**, An Introduction to Measure Theory, 978-7-04-046905-9
21. **Terence Tao**, Higher Order Fourier Analysis, 978-7-04-046909-7
22. **Terence Tao**, Poincaré's Legacies, Part I: pages from year two of a mathematical blog, 978-7-04-046995-0
23. **Terence Tao**, Poincaré's Legacies, Part II: pages from year two of a mathematical blog, 978-7-04-046996-7
24. **Cédric Villani**, Topics in Optimal Transportation, 978-7-04-046921-9
25. **R. J. Williams**, Introduction to the Mathematics of Finance, 978-7-04-046912-7

“度量几何”是建立在拓扑空间长度概念基础之上的处理几何的方法，这种方法在最近几十年飞速发展，并渗透到诸如群论、动力系统和偏微分方程等其他数学学科。

本书有两个目标：详细阐述长度空间理论中使用的基本概念和技巧，以及为大量不同的几何论题提供一个初等导引，这些论题都与距离观念相关，包括黎曼度量和Carnot-Carathéodory度量、双曲平面、距离－体积不等式、（大规模的、粗糙的）渐近几何、Gromov双曲空间、度量空间的收敛性以及Alexandrov空间（非正和非负的弯曲空间）。作者倾向于用“易于看见”的方法来处理“易于触碰”的数学对象。

作者设定了一个具有挑战性的目标，即让本书的核心部分能为一年级研究生所接受。大多数新的概念和方法都按最简单的情形来提出并阐明，从而避免了技术性的障碍。书中还包括大量习题，这些习题是本书至关重要的一部分。

本版只限于中华人民共和国境内发行。本版经由美国数学会授权仅在中华人民共和国境内销售，不得出口。

美国数学会经典影印系列



ISBN 978-7-04-046908-0



定价 169.00 元